6-2-2017

Automatically Guilty: Associations Between Evidence and Guilt

William E. Crozier IV
The Graduate Center, City University of New York
Automatically Guilty: Associations between Evidence and Guilt

William E. Crozier

John Jay College, CUNY
The Graduate Center, CUNY

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the Degree of Doctor of Philosophy, The City University of New York

2017
AUTOMATICALLY GUILTY

© 2017
WILLIAM EDWARD CROZIER IV
All Rights Reserved
AUTOMATICALLY GUILTY: ASSOCIATIONS BETWEEN EVIDENCE AND GUILT

WILLIAM E. CROZIER

This manuscript has been read and accepted for the Graduate Faculty in Psychology to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Deryn Strange
_________________
Date

Chair of Examining Committee

Richard Bodnar
_________________
Date

Executive Officer

Saul Kassin
Charles Stone
Christian Meissner
Stephan Dewhurst
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

AUTOMATICALLY GUILTY: ASSOCIATIONS BETWEEN EVIDENCE AND GUILT

by

WILLIAM EDWARD CROZIER

Advisor: Deryn Strange

Both real-life cases and laboratory research demonstrate that confession evidence is very convincing—even when it should not be. Could this be due to an automatic association between a confession and guilt? We tested this possibility using a Deese-Roediger-McDermott (DRM) list, which measures automatic associations by presenting participants with a list of words that are thematically related but, importantly, lack the word describing the theme (“critical lure”). When the association between the list words and the theme is sufficiently strong, participants incorrectly report seeing the critical lure. We hypothesized that participants would show more false recall for seeing “guilty” on a “guilty”-themed DRM list when the list included evidence that is automatically associated with guilt, such as “confession” and “DNA.” Although our previous research on this topic found no significant effects, we addressed limitations of that research in three studies using an Amazon Mechanical Turk sample. Our first study addressed a possible ceiling effect by decreasing the associative strength of our “guilty” list. Our second study increased external validity by

---

1 Because research is a collaborative endeavor, and this project would not have been possible without my advisor's support and contribution, I have chosen to use the pronoun “we” throughout my dissertation. I think it is the most accurate characterization of this work.
AUTOMATICALLY GUILTY

presenting our DRM List as a DRM Story—a narrative format that provides context for the list words. Our third study investigated the effects of priming evidence quality on the association to guilty.

Overall, we found little support for our hypotheses. Across all three studies, we did not detect any effects of the evidence type (Study 1, 2, and 3) or prime type (Study 3). We did, however, find several interesting trends in the data. We discuss explanations for the lack of significant findings and address directions for future research. Specifically, adapting this paradigm for other research applications and to increase our understanding of the memorial effects of the “guilty” DRM list.
Acknowledgements

First and foremost, I offer my eternal gratitude to my advisor, Dr. Deryn Strange. Without her constant patience, encouragement, support, and rapid email responding, I honestly believe this project would not have been possible. I have her to thank for the completion of my degree, as well as my insatiable desire to learn more about the world.

Second, I am similarly grateful to my partner, Lindsey Cartner, for her unwavering support during frustration. I can never overstate how essential she has been to the development of this project, every paper I've written, and my growth as scientist.

Third, my utmost appreciation to the members of my committee: Professors Saul Kassin, Charles Stone, Christian Meissner, and Stephen Dewhurst. Their genuine interest, critical questions, and enthusiastic suggestions in this project have improved this project and my skill as an experimental psychologist.

I also owe a debt to Dr. Timothy Luke, who proved a welcome resource for venting frustration, celebrating victories, and understanding the fundamentals of research. Whether in practical discussions about statistics or less-stressful conversations about unrelated research projects, his friendship and mentorship have been invaluable.

I would like to thank Ella Merriweather, Gabriella Rico, Caitlin McCauley, and Mashy Raul for their help in coding data and other miscellaneous contributions to the project.

Finally, I am appreciative to the Forensic Psychology Research Institute and, specifically Dr. Widom, for financially supporting my dissertation. The Dissertation Improvement Grant, and revisions required to receive the grant, both improved this project, and made it possible.
# Table of Contents

CHAPTER 1: THE CONNECTION BETWEEN EVIDENCE AND GUILT ........................................... 1

CHAPTER 2: TESTING FOR AUTOMATIC ASSOCIATIONS ............................................... 5
  - The Implicit Associations Test ................................................................................ 5
  - The Deese-Roediger-McDermott Paradigm ............................................................... 7
  - Activation-Monitoring Theory ................................................................................... 8
  - Using the DRM to Measure Automatic Semantic Associations .................................. 10

CHAPTER 3: DEVELOPING AND TESTING A "GUILTY" DRM LIST .................................. 14

CHAPTER 4: PROPOSED RESEARCH PLAN .................................................................. 19
  - Data Analysis Approach .......................................................................................... 21

CHAPTER 5: STUDY 1 – THE 7-WORD LIST ................................................................ 23

CHAPTER 6: STUDY 1A ............................................................................................. 24
  - Design .................................................................................................................... 24
  - Participants ............................................................................................................. 24
  - Materials ............................................................................................................... 25
  - Procedure ............................................................................................................... 25
  - Hypotheses ............................................................................................................ 26
  - Results ................................................................................................................... 27
  - Discussion .............................................................................................................. 29

CHAPTER 7: STUDY 1B ............................................................................................. 31
  - Design .................................................................................................................... 31
  - Participants ............................................................................................................. 31
  - Materials ............................................................................................................... 32
  - Procedure ............................................................................................................... 32
  - Hypotheses ............................................................................................................ 33
  - Results ................................................................................................................... 33
  - Discussion .............................................................................................................. 37

CHAPTER 8: STUDY 2 – THE DRM STORY FORMAT ..................................................... 40

CHAPTER 9: STUDY 2A ............................................................................................. 43
  - Design .................................................................................................................... 43
  - Participants ............................................................................................................. 43
  - Materials ............................................................................................................... 43
  - Procedure ............................................................................................................... 45
  - Results ................................................................................................................... 45
  - Discussion .............................................................................................................. 46
LIST OF TABLES

Table 1. Study 1b Remember/Know/Guess Results .................................................................37
Table 2. Study 2b Remember/Know/Guess Results .................................................................53
Table 3. Study 3a Sample 1 Evidence Prime Ratings ...............................................................65
Table 4. Study 3a Sample 2 Evidence Prime Ratings ...............................................................66
Table 5. Study 3a Sample 3 Evidence Prime Ratings ...............................................................67
Table 6. Study 3b Remember/Know/Guess Results ...............................................................77
LIST OF FIGURES

Figure 1. Study 1a Recognition Test Results .................................................................28
Figure 2. Study 1b Free Recall Results ...........................................................................34
Figure 3. Study 1b Recognition Test Results .................................................................35
Figure 4. Study 1b Confidence Results ...........................................................................36
Figure 5. Study 2b Free Recall Results ...........................................................................50
Figure 6. Study 2b Recognition Test Results .................................................................51
Figure 7. Study 2b Confidence Results ...........................................................................52
Figure 8. Study 3b Free Recall Results ...........................................................................73
Figure 9. Study 3b Recognition Test Results .................................................................75
Figure 10. Study 3b Confidence Results ........................................................................76
CHAPTER 1: THE CONNECTION BETWEEN EVIDENCE AND GUILT

At trial, a particular piece of evidence against a defendant typically suggests, but does not determine, guilt. Careful consideration should forge the association between evidence and guilt. That is, as a juror reasons through the facts of a case, they should be evaluating whether the evidence presented (both individual pieces and the body of evidence as a whole) is consistent with the defendant being guilty. But, what if the association between evidence and guilt is automatic rather than the result of careful decision-making? For example, if a juror simply hears “confession” do they automatically think and remember “guilty”?

Anecdotal evidence suggests that the association between confession and guilt may indeed be automatic. In its first 325 DNA exonerations, the Innocence Project found that 88 of these cases (27%) were due, in part or whole, to a false confession or false admission (www.innocenceproject.org/causes-wrongful-conviction). Thus, in this group of exonerations, someone involved in a criminal investigation—a detective, a district attorney, a prosecutor, a judge, a juror, or any combination therein—may have ignored potentially exonerating evidence and instead believed the confession. What about the confession—a form of evidence prone to error (Kassin et al., 2010)—made it so convincing, especially in the cases where there was contradictory DNA?

Consider a specific example: the Central Park Five case. On an April night in 1989, a female jogger was found brutally beaten and raped in Central Park. Police quickly rounded up a group of teenage boys seen roaming the Park earlier in the evening. Of those taken in for questioning, five (Yusef Salaam, Kevin Richardson, Antron McCray, Raymond Santana, and Korey Wise) confessed to the crime, saying the five of them assaulted the victim alone.
Eventually, police compared DNA recovered from a rape kit performed on the jogger to all five of the confessor's and found no matches (www.innocenceproject.org/cases-false-imprisonment/Korey-Wise). At this point, logic should dictate the teenagers' confessions and descriptions of the crime were all factually inaccurate, and thus the police should search for a new perpetrator. Instead, the district attorney charged the five confessors with the crime. And at trial, the jury made the same mistake: rather than question the lack of a DNA match, they were swayed by the confessions and found all five teenagers guilty.

Laboratory research has also demonstrated the power of a confession. Kassin and Neumann (1997) found mock jurors are most likely to convict when a confession is present, compared to cases with an incriminating eyewitness or an incriminating character statement. Furthermore, when mock jurors read a case containing all of these forms of evidence, they rated the confession as being the most incriminating. Problematically, research has also shown that confessions are persuasive—even when they should not be. To test whether a confession impacts a trial decision even if it should be ignored, Kassin and Sukel (1997) asked participants to read a trial transcript which contained no confession (control), or a confession from a fully-crossed 2(interrogation pressure: low, high) x 2(confession: admissible, inadmissible) design. If jurors could indeed ignore an improper confession in determining guilt, then only those jurors reading a case with a low-pressure admissible confession should convict at a higher rate than the no-confession control group. This hypothesis is based on the theory that a juror should understand an inadmissible confession or a confession produced from a high-pressure interrogation could lead to a false confession. Thus, a juror should ignore such a confession in their decision-making process. Instead, all of the participants who read a case with a confession were
more likely to convict than the control group. While the low-pressure admissible
confession group convicted the most (63% of participants), the low-pressure inadmissible
(50%) and high-pressure admissible (50%) and high-pressure inadmissible (46%) groups
were all significantly higher than the no-confession control (19%). Furthermore,
participants rated admissible confessions and low-pressure confessions as more influential
in their decision than the inadmissible- or high-pressure-counterparts. Thus, the
confession continued to affect participants’ decision making even when it should have been
ignored—and it did so without participants’ conscious awareness. Unfortunately, the
overriding power of a confession is not just limited to lay people. Recruiting 132 active
judges, Wallace and Kassin (2012) found that even they were vastly more likely to convict
if a confession was present— even when the interrogation resulting in the confession was
described as high-pressure, and despite the fact that the judges rated high-pressure
confessions as less voluntary.

In a recent review of confession and interrogation research, Kassin (2012)
investigated whether false confessions corrupt the truth-seeking process. He described
archival data that suggests defendants who confessed are more likely to have a bad defense
attorney (Kassin & Kukucka, 2012) and are more likely to plead guilty (Redlich, 2010) than
exonerees who had not confessed. With this body of research in mind, we wondered
whether it is possible that a confession could also corrupt the truth-finding process at an
individual, cognitive level. Perhaps the reason an improper confession is so difficult to
exclude from a juror’s guilt decision-making process is because a confession suggests guilt
so strongly that the confession is automatically and unconsciously associated with guilt —
AUTOMATICALLY GUILTY

and that automatic association is all but impossible to break. That is the proposition we address in this set of studies.
The Implicit Associations Test

When psychologists want to measure automatic associations, perhaps the most prominent paradigm that comes to mind is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). As its name implies, the IAT measures implicit associations by measuring how long it takes participants to categorize specific stimuli (i.e., African American names) into a proper category (Black names or White names) that is simultaneously paired with a seemingly unrelated category (Pleasant or Unpleasant). When participants take longer to categorize the stimuli into one classification (an African American name into the Black/Pleasant category) than another (African American names into the Black/Unpleasant category) they are said to have a preference for the shorter-response group. Although the IAT has been an incredibly fruitful paradigm, measuring a multitude of implicit attitudes and associations—such as race (Greenwald et al., 1998), gender (Greenwald, Nosek, & Banaji, 2003), voter behavior (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008), attitudes towards alcohol (Houben & Wiers, 2006) and smoking (Huijding, de Jong, Wiers, & Verkooijen, 2005), consumer behavior (Maison, Greenwald, & Bruin, 2004), gender and math (Nosek, Banaji, & Greenwald, 2002), and even as a questionably-effective tool to detect false memories (Sartori, Agosta, Zogmaister, Ferrerra, & Castiello, 2008; but see Takarangi & Strange, 2013)—we do not believe it is the best paradigm to test our question. The IAT has several theoretical problems and implication limitations that render it unsuitable for this research project. For example, although the IAT does reliably produce a difference in response times, it is still a matter of debate as to what that difference means. Indeed, Blanton and Jaccard (2006) refer to the
AUTOMATICALLY GUILTY

IAT as arbitrary, because it relies on (among other issues), millisecond metrics to measure bias, norming, and arbitrary zero points such that understanding differences in results requires a certain amount of interpretation. Moreover, Fiedler, Messner, and Bluemke (2006) found that the association model underpinning the IAT is questionable and, perhaps most relevant to this project, differences in IAT scores between individuals are ambiguous and hard to interpret (e.g., arbitrary zero points). These issues lead to a number of practical problems for this particular project. First, the IAT paradigm would be extremely difficult to modify. The IAT measures response times between two general concepts, such as “violent” and a racial group, averaged over many trials. The necessity of many repeated trials means that participants must view and respond to multiple iterations of the concept/association pairings—for example, twenty different photographs of African American males paired with ten “pleasant” and ten “unpleasant” stimuli. Because “guilty” is a specific concept that would be difficult to uniquely represent several times over, the IAT would be extremely difficult to adapt here; showing twenty African-American photographs is far easier than representing “guilty” in twenty different ways. Furthermore, it would require an extraordinarily large number of trials to compare guilty to different forms of evidence (such as confessions, DNA, etc.) to detect automatic associations. Second, the IAT would prove difficult to modify for follow-up studies as it is built in a very specific way and offers little flexibility in its construction and measurement. Finally, and perhaps most critical to our purposes, the IAT does not address memory. Although the IAT measures implicit associations, a paradigm that could also measure how people apply that association in a more directly-observable manner, would be much more useful.
AUTOMATICALLY GUILTY

The Deese-Roediger-McDermott Paradigm

The Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995) provides a novel way of testing potentially automatic associations in a legal context. Not only is it a much easier and more adaptable paradigm to implement (see below for its variety of research applications), it also relies on existing semantic knowledge networks and allows us to directly measure how people relate specific concepts through false recall formation. As such, we can observe how people relate certain forms of evidence to “guilty” as well as gain insight into how the resulting false memories may influence future decisions and judgments.

The methodology for the DRM paradigm is straightforward: participants view or hear a list of words that are related to a concept, or theme. However, the theme itself is not included in the list. This strongly-related, excluded word is referred to as the “critical lure.” For example, participants read: rest, doze, snore bed…, but not the critical lure “sleep.” Then, after a short delay to allow the list to pass out of short-term memory, researchers test participants’ memory for the list items, either by a) asking participants to recall all of the words they can remember in two minutes, or b) a recognition task for list words, new words, and the critical lure wherein participants rate whether a word is “Old” or “New,” and rate their confidence in their response. Furthermore, if “Old” is selected, participants categorize their memory as either a “remember” judgment (strong, vivid memory with specific details), a “know” memory (a general sense of familiarity), or a “guess” judgment. Research shows that participants will report having seen the critical lure, for a variety of list-themes, ranging from 27% (the king list) to 80% (sleep list) of the time on the recognition test (Stadler, Roediger, & McDermott, 1995). The phenomenological aspects of
AUTOMATICALLY GUILTY

the falsely recalled lure can also be similar to those for list words. For example, Read (1996) used the sleep DRM list and found that participants were just as likely to report having seen “sleep” on the list as they were any of the words actually presented. Overall, participants reported significantly different confidence ratings and “remember” judgments (signifying clear, strong memories) for the “sleep” critical lure and the list words. However, when participants estimated that the critical lure appeared in the first third of the list (compared to those that estimated the middle third or last third), confidence ratings and “remember” judgments were nearly identical for the critical lure and the list words.

**Activation-Monitoring Theory**

How does the DRM paradigm create such robust false recall? According to the activation-monitoring theory, people falsely recall the critical lure due to a combination of activating the concept, followed by committing a source monitoring error (Roediger & McDermott, 1995; Roediger & McDermott, 2000; Roediger, Balota, & Watson, 2001). As a person reads the list words, the topic to which they are all related is “activated” by one of two processes. According to the spreading activation account, the semantically-related list words cue the semantic network to which the critical lure belongs (Collins & Loftus, 1975). Subsequently, the critical lure is activated: Either the reader consciously thinks of the critical lure, as in “all of these words remind me of sleep” or without awareness, by which the reader thinks of the concept of sleep without realizing it. Alternatively, fuzzy-trace theory (Brainerd & Reyna, 1998) explains that as a person reads the list, they construct a gist representation of the list, which describes the properties and features common to all of the list words. The critical lure is also activated because it shares many of these features.
AUTOMATICALLY GUILTY

with the gist trace. However, as Gallo (2010) points out, these two processes are extremely similar, not mutually exclusive, and amount to the same end: activation of the critical lure.

Next, at test, the person must engage in a source monitoring decision to determine the source of the activated critical lure (e.g., sleep; Johnson, Hashtroudi, & Lindsay, 1993; Lindsay, 2008). When attempting to determine whether they have seen a specific word on the preceding list, people must access their memory for the list itself. When a word is not related to the list topic and was therefore not on the list (such as “lawyer” for the sleep list) it is easy to reject. However, when evaluating the source of the critical lure, which has been previously activated, a person must choose between an internal source (thinking about the word) and an external source (reading the word). When they falsely remember seeing the critical lure, they have committed a source monitoring error; they attribute their thinking of the critical lure to reading the word on the list.

At its heart, the DRM paradigm is a memory test. However, it relies on the strength of the semantic relationship between a critical lure and a topic — the stronger the relationship, the higher the false recall (or recognition) rate for the critical lure. Put another way, if the list words are sufficiently semantically related to the critical lure, that list will produce false recall. If, however, the list words do not relate to the critical lure, the list will fail to activate the network and participants will be less likely to falsely recall the word. Indeed, research has found a positive correlation between the associative strength of the list words to the critical lure, and the false recall of the critical lure: the more strongly associated the list, the more likely you are to see the DRM effect (Gallo, 2010; Roediger et al, 2001; Deese, 1959).
AUTOMATICALLY GUILTY

The DRM paradigm has been adapted in a number of ways to study memory more broadly, such as providing support for a dual-process memory model (Barnhardt, Choi, Gerkens, & Smith, 2006), studying development of memory in children (Metzger et al., 2008), memory in Alzheimer’s patients (Balota et al., 1999; Watson, Balota, & Sergent-Marshall, 2001), memory in PTSD patients (Brennen, Dybdahl, & Kapidzic, 2006; Hauschildt, Peters, Jelinek, & Moritz, 2012), and even measuring individual differences for false memory susceptibility (Blair, Lenton, & Hastie, 2002) — not to mention increasing our understanding of general true and false memory processes and how those processes develop (e.g., Roediger & McDermott, 1995; Roediger, Watson, McDermott, & Gallo, 2001; Dewhurst & Robinson, 2004; Barnhardt, Choi, Gerkens, & Smith, 2006; Brainerd, Yang, Reyna, Howe, & Mills, 2008; Sugrue, Strange, & Hayne, 2009).

Using the DRM to Measure Automatic Semantic Associations

The DRM paradigm can also be used to measure the semantic relationship between a list of words and the central concept due to its reliance on spreading activation within a semantic network. The more often a DRM list generates false recall for the critical lure, the greater the association between the lure and activated concept. If any alterations are made to the list words, any increase (or decrease) in the critical lure false recall rate is due to that alteration strengthening (or weakening) the semantic association. Similarly, any primes or stereotypes that activate related networks can also influence how often people false alarm to the critical lure. For example, Lenton, Blair, and Hastie (2001) found that participants were more likely to false alarm to a stereotypical male job or trait (i.e., engineer and wise) when viewing a list constructed of other stereotypical male jobs than when viewing a list of stereotypical female jobs. However, when viewing the stereotypical female jobs list,
AUTOMATICALLY GUILTY

participants false alarmed to female jobs and traits (dancer and delicate) more than stereotypical-male roles and traits. Furthermore, when participants were primed with a list of either male traits or female traits before viewing the male-job or female-job list, they showed greater false alarm rates when viewing the gender-associated list.

Takarangi, Polaschek, Hignett, and Garry (2007) also demonstrated that priming and individual differences can affect how we interpret a DRM list. In their study, when participants were shown an ambiguous list (words such as beat, batter, punch, etc. that could either be violent words or kitchen/food words), those who were primed to be more hostile, or were high in trait hostility, false alarmed to more non-ambiguous violent words (e.g., stab) than those low in state or trait hostility. Takarangi et al. explained that this higher false alarm rate was likely due to the strength of the violent-interpretation network being activated. Those who were more hostile were more likely to see the ambiguous list words in a hostile, violent way, and thus activated the unambiguously-violent critical lure. Therefore, because we know priming specific information can influence the semantic activation that leads to a DRM false alarm, we reasoned that it may be a useful tool to employ in the proposed studies. More specifically, could we introduce information about legitimate or wrongful convictions hinging on specific evidence via priming to strengthen or weaken that evidences’ association with guilt?

Because the associative strength of the list correlates with the critical lure, we should be able to use a DRM list to measure the associative relationship between different forms of evidence (the words on the list) and “guilty” as a critical lure. By creating a DRM list for the critical lure “guilty” and including different forms of evidence in the list, we should be able to use the rate of false recall for “guilty” to measure how closely each form of
evidence is associated with “guilty.” That is, if a piece of evidence, such as “confession,” is automatically strongly associated with guilty, then the semantic activation would be stronger than the basic list, thereby increasing false recall for the critical lure. Conversely, if a form of evidence is more weakly associated with guilty, the semantic activation would decrease compared to the basic list, and thereby generate less false recall.

Although confessions are an obvious candidate for evidence that might be automatically associated with guilt, there are certainly others. We suspect that evidence that is generally perceived to be highly reliable and highly accurate will be more likely to lead to a guilty verdict, and thus more strongly associated with guilt; the weaker the perception of the evidence, the weaker the association should be. For example, DNA is a form of evidence that we expect would be strongly associated with guilt, given that it has been scientifically developed and validated as reliable and accurate (albeit, not entirely free of error; Lieberman, Carrell, Miethe, & Krauss, 2008; National Academy of Sciences, 2009). Indeed, Lawson (2014) found that participants rated DNA evidence as being highly reliable (4.75 on a 1-5 Likert Scale). Importantly, participants rated DNA as more reliable than other forms of forensic evidence, including fingerprints (reliability = 4.33), bitemarks (3.89), toolmarks (3.58), and eyewitnesses (2.84) — a finding supported by Lieberman et al., (2008) and Hans, Kaye, Dann, Farley, and Albertson, (2011). Together, these ratings offer comparative forms of evidence for testing guilt associations. For example, eyewitness evidence is viewed as less reliable, likely because it is not scientific and is prone to error (National Academy of Sciences, 2014; Wells et al., 1998). As such, eyewitness evidence should be weakly associated with guilty and a good comparison for confessions which are similarly not the result of scientific analysis. On the other hand, fingerprint analysis is
AUTOMATICALLY GUILTY

similar to DNA in that it is scientific and perceived as fairly reliable, but, unlike DNA, not scientifically validated and thus more prone to error (National Academy of Sciences, 2009). Therefore, fingerprints should be associated with guilt, but not to the same extent as DNA. Likewise, bitemark analysis, despite its reputation, is not scientific, not validated, and prone to error (National Academy of Sciences, 2014) and has been the subject of significant negative media attention (Balko, 2015a, 2015b). Thus it should be weakly associated with guilt. Together, these forms of evidence represent a possible spectrum of guilt association, ranging from DNA (strong) to eyewitness (weak).
CHAPTER 3: DEVELOPING AND TESTING A “GUILTY” DRM LIST

In our initial tests employing the DRM paradigm to measure associations between guilt and evidence, we successfully created a 14-word “Guilty” DRM list. We had 60 participants from Amazon Mechanical Turk (MTurk) complete a forward associative task for the word “guilty.” They listed the first 10 words that came to mind when they read “guilty.” From these data, we chose the 14 most commonly reported words (excluding vague words, such as “bad;” See Appendix A)—which ranged from being listed 47% of the time (“jury”) to 10% of the time (“verdict” and “sentence”)—and tested whether a short (7 words) or long (14 words) list produced false recall of “guilty.” Forty-five participants read the list of words (presented for 1.5 seconds each), completed a 5-minute interference task, then completed an OLD or NEW recognition test and rated their confidence in each decision (1=not at all confident, 5=very confident). Participants were assigned to the 7-word list (N=20) or a 15-word list condition (N=25). All read the “sleep” list, the “chair” list (established DRM lists that we used as control comparisons) and the “guilty” list. We found that the word “guilty” was falsely recalled about 70% of the time, regardless of the list length (M7-word=.70; M14-word=.71), which is in line with our DRM control lists; “sleep” was falsely recalled 89% of the time (M7-word=.90; M14-word=.88); “chair” was falsely recalled 79% of the time (M7-word=.85; M14-word=.75). Thus, our “guilty” list reliably produced a high rate of false recall.

With a functioning “Guilty” DRM list, we then tested our primary research question: are different forms of evidence more strongly associated with guilt than others? To do so, we removed the last word (which is also the most weakly-associated word) and placed one of four types of evidence into the 7th position in the list, creating four new lists to test
AUTOMATICALLY GUILTY

between subjects. These forms of evidence were: “confession,” “DNA,” “bitemark,” and “eyewitness.” Participants were sorted into one of five conditions: a Control List condition, which read the basic “guilty” DRM list; a Confession List condition, which read the list containing “confession” as the 7th word; a DNA List condition; a Bitemark List condition; and finally an Eyewitness List condition. Based on the convincing nature of confession evidence (see Chapter 1), we hypothesized that participants in the Confession List condition would show significantly more false recall than all of the other conditions, save the DNA List condition. We further hypothesized that participants in the DNA List condition would show an increased false recall occurrence, similar to those in the Confession List condition, due to the fact that DNA evidence is scientifically reliable and widely used in court (National Academy of Sciences, 2009). Conversely, we expected the Bitemark List condition (evidence that is regarded as “scientific” but is not accepted by the National Academy of Sciences as reliable; National Academy of Sciences, 2009; Garrett & Neufeld, 2009) and the Eyewitness List condition (evidence that is prone to error and not regarded as scientific; National Academy of Sciences, 2014; Wells et al., 1998) to show fewer false recall for “guilty” than our strong-evidence conditions.

We recruited 699 participants from MTurk across three samples and assigned them to one of the five conditions (Control List N=146; Confession List N=151; DNA List N=159; Bitemark List N=148; and Eyewitness List N=95, smaller because we did not include it in our first sample). Surprisingly, we found no significant differences between groups ($X^2(4, N=699)=4.59, p=.332$; all pair-wise comparisons between groups were not significant, $p>.05$). However, we did find that participants reported reading “guilty” at consistently high rates: $M_{Control}=0.73, 95\%CI=[0.68,0.82]$; $M_{Confession}=0.82, 95\%CI=[0.75,0.88]$;
AUTOMATICALLY GUILTY

$M_{\text{DNA}}=0.82$, 95%CI=[0.76,0.87]; $M_{\text{Bitemark}}=0.85$, 95%CI=[0.79,0.90]; $M_{\text{Eyewitness}}=0.78$, 95%CI=[0.69,0.86]. Moreover, a one-way ANOVA on people's confidence ratings for reading “guilty” revealed that although the overall test was not significant ($p>.05$), participants in the Confession List group were more confident in their false recall of “guilty” than those in the Control List group ($t(132)=1.98$, $p=.049$, $d=0.258$; $M_{\text{Confession}}=4.35$, 95%CI=[4.20,4.45]; $M_{\text{Control}}=4.11$, 95%CI=[3.92,4.30]) and DNA List group ($t(152)=2.24$, $p=.026$, $d=0.283$; $M_{\text{Confession}}=4.35$, 95%CI=[4.20,4.45]; $M_{\text{DNA}}=4.09$, 95%CI=[3.92,4.26]).

Based on these data, we can conclude two things. First, manipulating the list to include specific forms of evidence did not seem to change the list’s activation strength of guilty. However, our results do seem to show an interesting trend. The lists containing evidence forms did generate more false alarms to “guilty” than the control list. Interestingly, the Eyewitness List did not generate as much false recall as the other evidence lists, likely due to its perceived weakness as a type of evidence (Lieberman, et al. 2008; Hans et al., 2011). The Bitemark List, however, produced high rates of false recall—rates in line with the Confession and DNA lists—which may be due to a lack of knowledge regarding the unreliability of bitemark evidence (National Academy of Sciences, 2009; Lawson, 2014). Second, although the types of evidence are not affecting how likely participants are to false alarm to guilty, there is some evidence that the variations in evidence may be affecting the phenomenological experience of the list. That is, people who read the Confession List were more confident that they saw guilty than participants who saw other forms of evidence. Despite the small effect size of these differences, we take this as evidence that there is some connection between confession and guilty compared to other forms of evidence.
Another reason to warrant further investigation with the “guilty” DRM list lies in the list’s pragmatic implication. As reviewed in Chapter 2, the DRM effect occurs when people (intentionally or unintentionally) connect all of the list words and semantically activate the theme—the critical lure. Our DRM list certainly seems to operate according to this mechanism. However, a body of research beyond the DRM paradigm provides insight into the false recall created by the “guilty” list. Work on pragmatic implication has demonstrated people can falsely remember implied, but not stated, ideas. Although this work has studied paragraph stimuli rather than lists, a similar mechanism may apply here. The criminal justice words on our list, such as “judge,” “jury,” “crime,” and “trial,” may pragmatically imply guilt causing people to falsely remember that implication as being a part of what they saw. Indeed, Bransford and Johnson (1971) found that participants would remember more information than was originally stated, or misattribute the source of that information, as a result of integrating the content of topically-related sentences.

Chan and McDermott (2006) note that the pragmatic implications of a message necessarily change the meaning of the original message. For example, participants who read “The baby stayed awake all night” later false alarmed to “The baby cried all night” 65% of the time. Of course, while babies do stay awake crying, they could be awake for a number of other reasons that do not require an upset infant. Further, they found that when a message was strongly implied—but not outright stated—participants remembered that message just as strongly as presented information. The implication here is that even though our list words should be guilt-neutral, people understand them to imply guilt. Indeed, such automatic association underlines the importance of further understanding this process. People should be innocent until proven guilty—not automatically guilty.
AUTOMATICALLY GUILTY

example, innocent and guilty people go before a judge, a jury, and have a trial. Further, the word “innocent”—the opposite of guilty—appears on the list. However, these words activate the justice system—a fixture of society that is both unique and familiar. Because the justice system’s courts are specifically designed to assess guilt, the unavoidable activation of guilt adds an additional level to the DRM semantic activation.

Given the mixed results from our previous work and the contextual richness of the concept of “guilty,” we designed a series of studies to further explore the usefulness of the DRM paradigm in detecting automatic associations between evidence and guilt.
CHAPTER 4: PROPOSED RESEARCH PLAN

Although our hypotheses were not generally supported in our preliminary work, the small effect we observed for Confession List confidence suggested this paradigm may still be useful. Indeed, there are at least two limitations we need to address. First, we suspected there may be a ceiling effect for our “guilty” DRM list. The Control List generates false recall for “guilty” in 73% of participants; a rate that places it with some of the most successful DRM lists in lure activation. Although not statistically different from Control, we see that the Confession List and the DNA List appeared to generate slightly higher rates of false recall (82% of participants for both lists). Although it may be that there is no difference in associative strength between the three lists and “guilty,” another explanation is plausible: the Control List association is already so strong that the addition of even more strongly associated words (“confession” and “DNA”) does little to increase the association.

Therefore, we addressed this possible ceiling effect in two ways. First, a shorter list could create a situation in which removal of the last word (the 7th word, compared to the 14th word in our previous study) would decrease the associative strength of the list more than our current 14-word list. Second, it is possible that some of our words were so strongly associated with guilt (the words at the beginning of the list) that the latter words did little to increase the association. Therefore, it may be possible to address the ceiling effect by using words that are more weakly associated with guilty, which would make the variations in guilt association between evidence types more detectable. Thus, a shorter list with more weakly associated words would be more sensitive to the associative strength of a single list word, thereby amplifying the associative strength of our evidence types. We test this hypothesis in Study 1.
A second important limitation to our prior work is the lack of context. Given the applied nature of our research question, we cannot ignore the possibility that situational context and previous knowledge would play a role in how a real juror would associate evidence with guilt. Reading a list of 7 or 14 criminal justice words is a far cry from an actual trial. Interestingly, Dewhurst, Pursglove, and Lewis (2007) demonstrated the DRM effect with a paragraph, rather than a list. The paragraph—termed a “DRM story”—placed the 14 list words in a narrative of 65-104 words, depending on the list. They developed this method to test children’s automatic associations. Children do not tend to show the DRM effect because they typically fail to make the semantic association between the list words (Brainerd, Reyna, & Forrest, 2002). The DRM story provides context for the list words and makes the semantic association easier for the children to detect. As a result, they are more likely to generate the critical lure than the original list approach. We adapted this “DRM Story” paradigm to our “guilty list” to create a more context-rich situation. Thus, we hypothesized that the increased context provided by the “DRM Story” would amplify the automatic association between “guilt” and strong evidence forms. We tested a “guilty” DRM Story format in Study 2.

Another important consideration regarding the issue of context is what people come into the study knowing about each form of evidence. So far, we have relied on participants’ existing knowledge-base. However, in a real case, the quality of the evidence plays a large role in how guilt is determined—good evidence (correctly gathered and analyzed) is far more likely to lead to a guilty verdict than bad evidence. Indeed, the quality of evidence (both general and for a specific piece of evidence in a specific case) is the type of contextual information that would be discussed in an actual case. Evidence quality information is also
AUTOMATICALLY GUILTY

akin to another real-world form of context—pretrial publicity (Studebaker & Penrod, 1997). Trial-relevant information learned outside of the courtroom (e.g., through media reports) can affect a variety of case-relevant judgments, such as defendant credibility and guilt ratings. Such effects are the result of a source monitoring error, in which people mistake what they learned pre-trial (which should be excluded from a juror’s consideration of the case) for being learned during the trial itself (Ruva & McEvoy, 2008). This is problematic for the chances of a fair trial, as only information learned during the trial is subject to the rules of evidence and should therefore be the only information used when determining a verdict. Although this project is not looking at pretrial publicity’s per se, there are clear parallels. Whereas pretrial publicity context provides additional case “facts,” priming evidence as should strengthen or weaken the relationship between evidence and guilt, thereby altering how people think about guilt.

As such, in Study 3 we addressed whether pre-list priming—to frame the quality of the evidence in the DRM list—strengthens or weakens the association between the evidence and guilt.

Data Analysis Approach

Because we are using the DRM to measure semantic activation, we are primarily interested in the differences between Evidence List groups. Thus, for the three studies we will report the free recall, recognition, confidence, and Remember/Know/Guess results in the main body. Other statistics of interest, but not experimental focus, are included in the Appendices. For example, the memory measures for old words, new words, and the weak lure for the “guilty” list and comparison DRM lists (“chair” for Study 1; “sleep” for Study 2) are included in the Supplementary Statistics Appendices for the respective studies.
Across all three experiments, Free Recall and Recognition Test results are submitted to ANOVAs. Because these measures are dichotomous (e.g., “old” or “new”), these data are traditionally submitted to chi-square analyses. We chose to report ANOVAs and t-tests for two reasons. First, chi-square analyses yield the same significance values as ANOVAs and t-tests. To this point, we analyzed our data both ways and include the chi-square results in the Appendices for the respective study. Second, ANOVAs and t-tests are not entirely inappropriate statistical tests, as we are attempting to compare the rates of false alarms per group—which is a continuous variable. Indeed, we find that ANOVA results and effect sizes are more intuitively understood than chi-square results and effect sizes.
CHAPTER 5: STUDY 1 – THE 7-WORD LIST

An alternate explanation of our original findings is that our Guilty list was so strong that we observed a ceiling effect in the rate of critical lure false recall. The goal of this study was to test the automatic associations between evidence and guilt with a control list that is more weakly associated with “guilty.” To accomplish this, we decreased the list from a 14-word list to a 7-word list and address the question: Do some forms of evidence show greater automatic associations with guilt than others in a 7-word list? Traditionally, a DRM list presents words in order of decreasing association: thus, the strongest-associated word is presented first, and the weakest-associated word is presented last (Roediger & McDermott, 1995; Stadler, Roediger, & McDermott, 1999). However, because it was our goal to decrease the associative strength of our list, including the most strongly-associated words may not be the best strategy. Therefore, which words we ultimately include in our 7-word list is as important a consideration as the length of the list. We hypothesized that shortening the list to 7-words by removing the three most strongly associated-words and the 4 most weakly-associated words, false alarm rates to the critical lure will decrease (Study 1a). Then, using this new 7-word list, we hypothesized participants in a Confession List and DNA List condition would false alarm to the critical lure more than participants in the Control List, Fingerprint List, and Eyewitness List conditions (Study 1b).
CHAPTER 6: STUDY 1A

Design:

Between-subjects 3 group design (Relative Forward Associative Strength of the Words: Words 1-7, strong association: 47%-24%; Words 4-10, medium association, 39%-17%; Words 7-14, low association, 24%-10%; see Appendix A for List Words and their Forward Associative Strength, and Appendix C for the word compositions of the lists). Because the results of our first sample were surprising, we gathered a second sample to replicate the findings. In this second sample, we included a fourth word list (a Medium-Strong Association list, comprised of words 2-8, 44%-24% BAS), creating a 4 group design (BAS of words: Strong Association; Medium-Strong Association; Medium Association; Weak Association).

Participants:

Sample 1. We recruited 201 MTurk workers. Of these, we excluded 53 for failing to follow directions (e.g., taking notes, engaging in another task, watching the word list (which plays as a movie) more than once) and/or technical issues (e.g., the list failing to load, Qualtrics crashing). For full exclusion criteria, see Appendix B. Thus, our final sample included data from 148 participants (52 saw the Strong Association list; 44 the Medium Association list; 52 the Weak Association list; imbalance due to exclusions). Participants were 96 female, 52 male, with an average age of 37.38 (SD=11.19). English was the primary language of all participants. Twenty-six held graduate degrees, 68 college degrees, and 54 high school diplomas.

Sample 2. We collected 251 MTurk workers. Of these, 50 were excluded, leaving 201 in our final analysis (46 Strong Association list; 52 Medium-Strong Association list; 53
AUTOMATICALLY GUILTY

Medium Association list; 50 Weak Association list). Participants were 137 female, 64 male, with an average age of 37.93 (SD=11.68). English was the primary language of all participants. Twenty of the participants held graduate degrees, 90 a college degree, and 91 a high school diploma.

We hypothesized that weakening the associative strength of the list would have a medium-sized effect on false alarm rates ($d=0.5$). Thus, we needed approximately 50 participants per cell to detect differences in the critical lure false alarm rate between lists with 80% power and alpha=0.05.

Materials:

**DRM Lists.** Participants saw two 7-word DRM lists in a video: The “chair” list and the “guilty” list, manipulated for forward associative strength (see Appendix A). Because the “chair” list seemed to have rates comparable to our “guilty” list in pilot testing, we chose to use this as a DRM memory control. We excluded the “sleep” list (that we have used in past studies) because the false alarm rate is extremely high (and thus a weaker control than the “chair” list) and a second control DRM list adds unnecessary time to the experiment.

The four lists we used (3 in Sample 1, all 4 in Sample 2) were derived from our previous development of the “guilty” DRM list. The lists contained seven words from our previous 14-word list. For a list of the word lists used in Study 1a, see Appendix C. For the instructions and construction of the memory measures, see Appendix L.

Procedure:

Participants were recruited via MTurk and directed to Qualtrics.com to complete the survey. Participants completed two DRM lists (“chair” and “guilty”) with the following components: first, participants watched a video presenting the 7-word list with each word
AUTOMATICALLY GUILTY

presented for 1.5 seconds. Then, after a five-minute interference task, they entered as many words as they could remember from the list in 2 minutes, our free recall measure. Next, they completed a recognition test measure. It included words on the list (list words), words not on the list (new words), and the unifying theme word (the critical lure: either chair or guilty). For each of these words, participants responded whether the word was “Old” (on the list) or “New” (not on the list) and rated their confidence for this decision on a 1-5 Likert scale. Finally, if participants rated the word as “Old” they were asked to categorize their memory as “Remember” (the memory for the word is clear, with specific details), “Know” (the participant does not remember specific details but has a general sense of familiarity), or “Guess.” See Appendix L for the memory tests.

After watching the video and completing the memory tests for the “chair” list, participants repeated this procedure for the “guilty” list. The “chair” and “guilty” lists participants viewed were determined by random assignment, with participants placed into the Strong Association group, Medium Association group, or Weak Association group (between subjects; in Sample 2, a quarter of participants were placed into a Medium-Strong Association group; Appendix C).

Hypotheses:

$H_1$: The association strength of the “guilty” list will generate differences in false alarm rates (both free recall and recognition). The strong association “guilty” list will produce the most false alarms to “guilty,” followed by the medium association list, and lastly the weak association list.

$H_2$: The association strength of the “guilty” list will generate differences in confidence rates for false alarms. The Strong Association “guilty” list will produce highest
AUTOMATICALLY GUILTY

confidence for false alarms to “guilty,” followed by the Medium Association list, and lastly the Weak Association list.

$H_3$: The association strength of the “guilty” list will generate differences in remember/know/guess for false alarms. The Strong Association “guilty” list will produce the most “remember” judgments for “guilty,” followed by the Medium Association list, and lastly the Weak Association list. Conversely, the Weak Association list will produce the most “guess” judgments for “guilty,” followed by the Medium Association list and Strong Association list.

Results:

Recognition Judgments. In Sample 1, the words that made up the “guilty” list did have a significant impact on false recall rate for the critical lure, $F(2, 145)=3.99$, $p=.021$, $\eta^2_p=0.052$. The forward associative strength of the lists, however, had the opposite effect than we predicted: Participants in the Strong Association list had the lowest false alarm rate to “guilty” ($M_{\text{Strong}}=0.48$, 95% CI=[0.34,0.62]), whereas participants in the Weak Association List had the highest false alarm rate ($M_{\text{Weak}}=0.73$, 95% CI=[0.61,0.86]) with participants in the Medium Association list in the middle ($M_{\text{Medium}}=0.68$, 95% CI=[0.54,0.83]). This trend was not consistent with the “chair” list control, where the Strong Association list generated the most false alarms ($M_{\text{Strong}}=0.79$, 95% CI=[0.67,0.90]) compared to the Medium ($M_{\text{Medium}}=0.59$, 95% CI=[0.44,0.74]) and Weak ($M_{\text{Weak}}=0.63$, 95% CI=[0.50,0.77]). Thus, it seems that our “guilty” list is behaving differently from the “chair” list in the predictive value of the forward associative strength on the DRM effect. Hence, we gathered a second sample in order to replicate this trend.
Sample 2 produced the same general trend as Sample 1. Recall that we introduced a new list (words 2-8, the "Medium-Strong" list). The Medium-Strong list produced the lowest rate of false recall ($M_{\text{Medium-Strong}}=0.44$, 95% CI=[0.30,0.58]), followed by false alarm rates similar to Sample 1 for the other lists: the Weak Association list with the most false alarms ($M_{\text{Weak}}=0.70$, 95% CI=[0.57,0.83]), followed by the Medium Association ($M_{\text{Medium}}=0.57$, 95% CI=[0.43,0.70]), and the Strong Association list ($M_{\text{Strong}}=0.50$, 95% CI=[0.35,0.65]). There was, however, no main effect of the list type ($F(3, 197)=2.55$, $p=.057$, likely due to the minimal difference between the Medium-Strong and Strong lists.

**Figure 1.** Study 1a Proportion of “Old” responses on Recognition Test

**Confidence Judgments.** We submitted Confidence judgments for the “guilty” critical lure to a one-way ANOVA. Sample 1 data showed a significant main effect of list strength, $F(2, 145)=6.14$, $p=.003$, $\eta^2_p=0.04$. This main effect was driven by the fact that participants who saw the Strong Association list were significantly more confident in their false alarms ($M_{\text{Strong}}=4.13$, 95% CI=[3.79,4.48]) than were participants who saw the Medium Association
AUTOMATICALLY GUILTY

list ($M_{Medium}=3.61$, 95% CI=[3.22,4.00]; $t(94)=2.01$, $p=0.047$, $d=0.415$, 95% CI=[0.17,0.67]) or participants who saw the Weak Association list ($M_{Weak}=3.58$, 95% CI=[3.21,3.94]; $t(102)=2.224$, $p=0.028$, $d=0.43$, 95% CI=[0.19,0.68]).

This finding did not replicate in Sample 2, however, because there was no main effect for list type on confidence ($p=0.809$), and there were no significant differences in confidence for participants in the Strong Association list ($M_{Strong}=3.61$, 95% CI=[3.28,3.94]) and any other lists.

*Remember/Know/Guess Judgments.* We submitted the R/K/G judgments to a chi-square test. For both Sample 1 and Sample 2, the chi square tests were not significant (Sample 1: $X^2(4, N=93)=1.87$, $p=0.759$; Sample 2: $X^2(6, N=111)=5.14$, $p=0.526$). Follow-up comparisons revealed no significant differences between any of the specific conditions in both samples (all $p$’s>0.05). Generally speaking, participants tended to make approximately the same number of “remember” or “know” judgments per evidence list, while slightly fewer made “guess” judgments. Thus, list strength had no apparently effect on the R/K/G judgments for “guilty.”

**Discussion:**

Recall that the goal of Study 1a was to determine which 7-word list would best decrease the false alarm rate of the “guilty” list to a rate that could avoid a ceiling effect and thus allow for more upward movement in our planned studies. Given that our primary variable of interest is the Recognition Judgment, and neither Confidence Judgments nor R/K/G Judgments revealed any notable trends, we drew our conclusions for this study based primarily on the Recognition false alarm rates. Although we predicted that the Medium or Weak list may be best suited, it seems the Strong list reliably produced a false
alarm rate low enough to avoid a ceiling effect. Although the Medium-Strong also showed a lower activation rate, we only included it as an extra measurement point when replicating Sample 1. Thus, the Medium-Strong list may be useful in future research, but based on two samples, the Strong list is the best candidate for testing the main research question of Study 1b: Can we use the 7-word list to measure differences in associations between guilty and forms of evidence?
CHAPTER 7: STUDY 1B

Design:

Between-subjects 5 group design (Evidence form: Control, Confession, DNA, Fingerprint, Eyewitness). In a departure from our earlier work, we used fingerprint as a form of evidence instead of bitemark. Participants in the Bitemark List condition did not show different effects from the DNA list. Furthermore, fingerprints are similar to confessions in a way DNA and eyewitnesses are not. As evidence, DNA or eyewitnesses can be offered as incriminating or exonerating. However, confessions are only incriminating. Similarly, fingerprint matches are offered primarily as incriminating evidence. Thus, a Fingerprint List condition allows us to compare the association between guilt and evidence that can indicate guilt or innocence to evidence that only indicates guilt.

Participants:

We collected data from 1423 MTurk workers. Of these, we excluded 407 for failing to follow directions and/or technical issues. Thus, our final sample included 1016 participants (206 saw the Control list; 200 the Confession list; 204 the DNA list; 214 the Eyewitness list; and 192 the Fingerprint list). Participants were primarily female (N=714; male N=302), with an average age of 37.53 (SD=11.97). Given that our previous findings had effect sizes that were small to insignificant, we assumed differences in evidence inclusion would produce small-to-medium effect sizes. Thus, a priori power analyses calculate that this sample size, approximately 200 per cell, will be sufficient to detect a small-to-medium effect size ($d=0.30$) with 80% power.

English was the primary language of all but 16 participants; however these 16 participants provided free responses in the free recall sections and open-ended comments.
sections without spelling errors and with good grammar. Further, excluding these participants did not change the pattern of results. One hundred and twenty-eight of the participants had earned a graduate degree; 430 earned a college degree; 545 earned a high school diploma; and 4 had not finished high school.

We also asked participants whether they had served on a jury (Yes=181 participants), and whether a close friend or family member had ever been convicted of a crime (Yes=227 participants). Excluding these participants did not change our results for any of the analyses, and did not differ from participants who responded “no,” (with the exception of Confidence for participants who had served on a jury; this subset is reported below in the appropriate Results section) so we report data with all participants included.

**Materials:**

*DRM Guilty List.* Based on the results from Study 1a, we used the Strong Association 7-word “guilty” DRM list as our base Control list. Then, as in previous work, we removed the 7th word from the list and inserted the experimentally-manipulated form of evidence into fourth position on the list. Thus, we created five 7-word lists: a Control list, a Confession list, a DNA list, an Eyewitness list, and a Fingerprint list (see Appendix C for the full lists).

*DRM Chair List.* For a control, we used the first seven words in the “chair” DRM list (Roediger & McDermott, 1995).

For the instructions and construction of the memory measures, see Appendix L.

**Procedure:**

Recruitment, list presentation, and memory measures were identical to Study 1a. Participants first viewed a “chair” list, completed memory measures, and then viewed one of five “guilty” lists, based on random assignment: Control, Confession, DNA, Eyewitness,
AUTOMATICALLY GUILTY

*Fingerprint.* After viewing the “guilty” list and a five-minute delay task, participants completed memory measures (free recall, recognition old/new judgments, confidence judgments, and remember/know/guess judgments if they judged a word as “old”) for the list, and were then debriefed.

**Hypotheses:**

*H₄:* The *Confession* and *DNA* lists will generate the most false alarms, while the *Control* list will generate the least false alarms. The *Eyewitness* and *Fingerprint* lists will generate fewer than the *Confession* and *DNA* lists, but more than the *Control*.

*H₅:* The *Confession* and *DNA* lists will generate the highest confidence for the critical lure of the lists, followed by the *Eyewitness* and *Fingerprint* lists. The *Control* list will generate the lowest confidence for the critical lure.

*H₆:* The *Confession* and *DNA* lists will generate the most “remember” judgments and fewest “guess” judgments for false alarms, while the *Control* list will generate the fewest “remember” judgments and most “guess” judgments. The *Eyewitness* and *Fingerprint* lists will generate intermediary “remember” and “guess” judgments.

Based on our previous work, we expect these differences between groups to be small effect sizes (Cohen’s *d*=0.2)

**Results:**

*Free Response.* Notably, the false recall rates for each group were extraordinarily low, ranging from $M_{DNA} = 0.05$, 95% CI=[0.02, 0.04] to $M_{Control} = 0.09$, 95% CI=[0.05, 0.12]. We submitted these data to a five-way ANOVA, which was not statistically significant, $F(4, 1011) = 0.77$, $p = .544$. All follow-up t-test comparisons were insignificant as well, $p = [.088,
AUTOMATICALLY GUILTY

.904]. Thus, our hypothesis was not supported: there were no significant differences in activation of “guilty” based on the evidence in the list.

Figure 2. Study 1b proportion of “guilty” responses to Free Recall Test

Recognition Judgments. In line with Study 1a results, false alarm rates to the critical lure were much lower than our previous work (participants who read the Control list had a mean false alarm rate of $M_{Control}=0.48$ 95% CI=[0.41, 0.55]). However, there was little variation between the all of the lists (minimum: $M_{DNA}=0.46$ 95% CI=[0.39, 0.52]; maximum: $M_{Confession}=0.53$ 95% CI=[0.46, 0.60]. Statistical tests supported the lack of difference between groups, as the five-level ANOVA was not significant, $(F(4, 1011)=0.68, p=.603)$ and none of the between-group follow-up $t$-tests were significant ($p=[.137, .983]$). Despite not achieving statistically significance, it is interesting that the Confession List yielded the highest false alarm rate, while the DNA List yielded the lowest. Although our 7-word list successfully decreased the average false alarm rate for the critical lure, the results do not
support our hypothesis of the DNA and Confession Lists causing more false alarms than the other lists.

Figure 3. Study 1b proportion of “Old” responses on Recognition Test

Confidence Judgments. Confidence ratings (1-5 Likert Scale; higher scores indicating greater confidence in Old/New judgments) were submitted to a five-level ANOVA, and similarly yielded no significant differences, $F(4, 1011) = 1.23, p = .296$. Indeed, no follow-up $t$-tests between groups were significant either: maximum $M_{\text{Control}} = 3.83$, 95% CI=[3.67, 3.99]; minimum $M_{\text{Confession}} = 3.60$, 95% CI=[3.43, 3.77]; $p = [.052, .955]$. Thus, these data provided no support for our hypothesis that the DNA and Confession list would produce higher confidence for “guilty” recognition judgments than the other lists.
As mentioned previously, participants who had served on a jury \((N=178; \text{between 27 and 46 participants per group})\) showed significant differences in confidence ratings for each list, \(F(4, 177)=2.71, p=.032\), motivated primarily by significantly higher confidence for the DNA List \((M=4.11, 95\% \text{ CI } [3.88, 4.34])\) than the Confession List \((M=3.36, 95\% \text{ CI } [2.92, 3.81]); t(80)=3.23, p=.002, d=0.729 95\% \text{ CI } [0.506, 0.952]). However, because we find no other significant effects for responses from this subgroup, and this specific effect disappears from the overall sample, we refrain from interpreting this result.

**Remember/Know/Guess Judgments.** We submitted R/K/G judgments to a chi square test of independence to investigate whether the evidence form on the list impacted the phenomenological quality of false alarming to “guilty.” Recall that we only asked those who selected “old” to make a R/K/G judgment. We found no support for our hypothesis, the chi square was not significant \(\chi^2(8, N=505)=6.86, p=.552\). Follow-up comparisons revealed no
significant differences between any of the specific lists (all p’s>0.05). Generally speaking, participants tended to make approximately the same number of “remember” or “know” judgments per evidence list, while slightly fewer made “guess” judgments (See Table 1).

Table 1

<table>
<thead>
<tr>
<th>List Type</th>
<th>Remember</th>
<th>Know</th>
<th>Guess</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>35</td>
<td>39</td>
<td>25</td>
<td>99</td>
</tr>
<tr>
<td>Confession</td>
<td>34</td>
<td>41</td>
<td>31</td>
<td>106</td>
</tr>
<tr>
<td>DNA</td>
<td>38</td>
<td>35</td>
<td>20</td>
<td>93</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>32</td>
<td>31</td>
<td>32</td>
<td>98</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>35</td>
<td>31</td>
<td>32</td>
<td>98</td>
</tr>
</tbody>
</table>

Discussion:

Recall that the goal of the first study was twofold: first, Study 1a aimed to use a 7-word list that reduced the rate of false recognition for the critical lure “guilty,” and Study 1b tested for automatic associations between specific forms of evidence and guilt by placing the evidence as one word in the shorter, weaker “guilty” list.

Study 1a was, overall, successful. Participants falsely remembered seeing “guilty” on the 7-word list about 50% of the time, thus providing us with a usable 7-word list for Study 1b. However, this success comes with a surprising trend. We expected forward associative strength to be positively correlated with the rate of false alarms to “guilty,” such that the Strong and Medium-Strong lists generate the most false alarms and the Weak list to generate the least. In fact, we found the opposite: Weak (Sample 1 M_{Weak}=0.73; Sample 2 M_{Weak}=0.70) and Medium (Sample 1 M_{Medium}=0.68; Sample 2 M_{Medium}=0.57) lists generated the highest rates of “guilty” false alarms—in some cases, rates similar to the 14-word Control lists in previous studies. The Strong list, on the other hand, generated a false alarm
rate closer to the 50% we desired (Sample 1 $M_{\text{Strong}}=0.48$; Sample 2 $M_{\text{Strong}}=0.50$). This negative correlation is surprising in light of previous research that forward associative strength is positively correlated with critical lure false alarm rates (Gallo, 2010). Indeed, we found this positive correlation with our control “chair” list. Thus, it seems that for our list, a word’s forward associative strength to “guilty” may not be directly relatable to its backward associative strength with “guilty” (which we have not measured). We believe this is a feature of the “guilty” list deserving of future investigation. Indeed, a first step would be to measure the BAS between the list words and “guilty” and compare those to our FAS values from previous work. By doing this, we may verify that the backward associative strength is, in fact, stronger for our Weak list, and hence explaining the surprising results of Study 1a.

In Study 1B, we expected that a weaker, shorter 7-word list would make this paradigm more sensitive than our previous work to associations between forensic evidence and guilt in two ways. First, we expected that the lower false alarm rate would avoid a potential ceiling effect by allowing more “room” for a strongly-associated word to activate the critical lure at a higher rate. Second, we expected that fewer words would allow each word to proportionately contribute more to activating the critical lure, thereby magnifying the effect of changing a single word (the form of evidence).

Despite our weakened list, we did not find support for our hypotheses that lists containing DNA and Confession would generate more false alarms to “guilty” than other lists across all four of our dependent measures. One possible explanation for the lack of significant differences is that even with a 7-word list, changing a single word on the list is insufficient to alter the semantic activation strength for “guilty.” Indeed, as the results of
AUTOMATICALLY GUILTY

Study 1a show, cutting the list in half only reduced false alarm rates by approximately 5%-25%, compared to previous work with the 14-word list. Thus, while the 7-word list may avoid a ceiling effect in false alarm rate, it may not prevent a ceiling effect for the proportional contribution of each word in activating the semantic network. Successful development of this paradigm may require a list made of words with very weak forward associative strength with “guilty”—metrics that we do not presently have and, given the seemingly negative relationship between FAS and false alarm rates (Study 1a), we cannot estimate.

In summary, using the 7-word list did not yield different results from a 14-word list. Although the false recognition rates were in line with what we suspected would make for a more sensitive paradigm, we failed to find any significant differences between the different evidence lists. Thus, while these results are interesting from a DRM-design standpoint, a traditional DRM list appears ineffective for measuring automatic associations between evidence and guilt. Fortunately, we have other ecologically-valid methods of influencing the semantic activation of the critical lure—namely, by manipulating the context in which the list words are given. Our next step was to test for automatic associations with a more context-rich paradigm.
CHAPTER 8: STUDY 2 – THE DRM STORY FORMAT

Given the contrived nature of a list of words used in the previous experiment, our next goal was to increase the ecological validity of our guilty associations paradigm to a method more in line with how people normally encounter information. One way to accomplish this goal is to present the DRM words in a story structure, allowing us to test whether some forms of evidence show greater association with guilt than others in a more context-rich paradigm. By presenting a DRM Story instead of a DRM list, we hypothesized the context provided by the Story format would increase the association between the list words and the critical lure (Dewhurst, Pursglove, & Lewis, 2007). More specifically, instead of having the “guilty” list words presented one a time, here we instructed participants to read a paragraph containing these words. For example, participants read: “A jury is a group of citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order...”

The story format provides context that, by explaining how a trial works, makes the connection between the words easier to detect. We expected that adding this context would produce a deeper semantic processing of the words, a process that, compared to shallow processing, produces more false alarms to the critical lure (Toglia, Neuschatz, & Goodwin, 1999; Thapar & McDermott, 2001). It is likely that when people read about information for a criminal investigation, they automatically consider the evidence present to determine guilt. That participants may evaluate the story (focusing on evaluating evidence) is, to our knowledge, unique to our DRM list and should result in a deeper semantic connection to “guilty.” In turn, the more strongly a piece of evidence indicates guilt, the stronger the activation of the concept of guilt. As a result, we expected that
AUTOMATICALLY GUILTY

participants in the *Confession* Story and *DNA* Story conditions would false alarm to the critical lure more than participants in the other groups, enhanced by the context and deeper semantic processing of the evidence in the Story.

The research on false memories for implied content is particularly relevant here. Chan and McDermott (2006) found that people can form false memories for seeing implied, but not presented, information. To elaborate, they asked participants to read sentences that pragmatically implied, but did not explicitly state, some idea. Later, the researchers tested the participants’ memory for the sentences they read, and sentences the participants did not read but contained the implied message. For example, participants would read “The new baby stayed awake all night” but then see at test “The new baby cried all night” and rate their memory for the sentence with a “remember”, “know”, or “guess” judgment. Chan and McDermott found that participants generally responded to the implied messages with “remember” and “know” at rates very similar to presented information. The paradigm Chan and McDermott used is analogous to what we designed here. It stands to reason that many participants could read a DRM story about “guilty” and derive the implication the paragraph describes a guilty person. Although our memory test is focused specifically on single words, rather than whole sentences and the message they communicate, we may still be able to elicit false alarms to “guilty” with a DRM story. In fact, given Chan and McDermott’s findings that implied information was recalled with the same phenomenological strength as studied information, the added context may strengthen the memory quality for “guilty” in addition to increasing the frequency of false alarms.

Alternatively, it is possible that a DRM Story would actually *decrease* false alarm rates to the critical lure. Previous research has found that ease of theme identifiability is an
important predictor for false recall of the critical lure: When a theme is highly identifiable, (as in a DRM Story, the very reason it was developed to be used with child participants), adults show a decreased false alarm rate for the critical lure (Carneiro, Fernandez, & Dias, 2009). Carneiro et al. (2012) argued that this seemingly counter-intuitive finding is due to adults’ use of an “identify-to-reject” strategy: if a participant is able to identify the theme of a list (the critical lure) and notice that specific word is missing, they will not show the DRM effect. Therefore, if our DRM Story makes the theme of “guilty” highly identifiable, it could lead to a decrease in false alarms. Furthermore, if this increase in theme identifiability occurs, we could have an effect opposite of what we predict for our evidence forms: by removing an associated word and including a weakly-associated type of evidence, “guilty” could be less identifiable, thereby increasing false alarms for the critical lure.

As with Study 1, Study 2 was conducted in two parts: In 2a, we developed a “guilty” DRM Story format that we could use for 2b. In 2b, we used that “guilty” DRM Story to test automatic associations between evidence and guilt in a manner similar to Study 1b. As such, we can make the same predictions regarding false alarms to “guilty” here as in Study 1. We expect participants reading a “guilty” story that contains DNA and confession to false alarm to the critical lure more often than participants who view a story containing more weakly-associated forms of evidence or those viewing the Control story.
CHAPTER 9: STUDY 2A

Design:

There were no experimental manipulations in this study. In two samples, all participants read the same “guilty” DRM story. Each sample viewed a slightly different version of the “guilty” DRM story (Sample 1: 14-word; Sample 2: 7-word)

Participants:

Sample 1. We collected data from 75 MTurk workers. Of these, we excluded 6 for failing to follow directions and/or technical issues. Thus, our final sample included 69 participants. There were 27 females, 25 males, and they had an average age of 37.99 (SD=12.02). English was the primary language of all participants. Eleven held graduate degrees, 30 college degrees, and 27 high school diplomas, and 1 did not complete high school. ²

Sample 2. We collected 31 data from MTurk workers. Of these, 7 were excluded, leaving 24 in our final analysis. There were 14 females, 10 males, and they had an average age of 38.79 (SD=10.63). English was the primary language of all but one participant. Two held graduate degrees, 11 college degrees, and 8 high school diplomas, and 2 did not finish high school.

Materials:

We adapted our “guilty” DRM list to construct two “guilty” DRM Stories: a 14-word and 7-word version. Both Stories generally describe the legal process. The goal was to create a list that produces 30% recognition false alarms to the critical lure. We proposed

² This is the same sample as Study 3a Sample 1. After reading the Story Format and taking the memory test, these participants then read the evidence primes in Study 3a.
AUTOMATICALLY GUILTY

30% as a cut off because it conforms to previously-developed DRM lists ("king", 27%; Stadler, Roediger, & McDermott, 1995), but we expected to have a lower false alarm rate than the list-format in Study 1 because our adults may use a “search-to-reject” strategy, minimizing the DRM effect.

The 14-word version contained the fourteen “guilty” list words from previous work, and was 133 words overall (emphasis added). The 14 words appeared in the same order as the list, that is, ordered from strongest to weakest forward associative strength.

14-word Story (Sample 1) A jury is a group of normal citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order and gives instructions. In some cases, family members and interested community members can be present in the court as well. If a person loses, they will be considered a criminal. If they win, they are usually innocent and show this by the end of the trial. If they lose, they are sent to jail, especially for murder. They are also labeled as a convict, which sometimes makes them feel shame. Of course, a person is not alone – they usually have a lawyer on their side to argue before the verdict, as well as help them get a fair sentence. Thus, many people are involved in the system.

The 7-word version contained the same seven words used in Study 1b, and was 72 words overall. We tested a 7-word paragraph for the same reason we developed a 7-word list: a 7-word paragraph has a better chance of avoiding a false alarm ceiling rate, and the substitution of one word should have a larger impact on semantic activation.

7-word Story (Sample 2) A jury is a group of normal citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order and gives instructions. In some cases, family members and interested community members can be present in the court as well. If a person loses, they will be considered a criminal. If they win, they are usually innocent and show this by the end of the trial.

For the instructions and construction of the memory measures, see Appendix L.
Procedure:

We directed our mTurk participants to Qualtrics.com where they viewed the “guilty” DRM Story. After a five minute delay, participants completed a Free Recall measure, where we asked them to replicate the story paragraph they read as closely as possible. Afterward, they completed the same Recognition test (Old/New judgments), Confidence measure, and Remember/Know/Guess judgments for words they determined were Old, as in Study 1a and 1b.

Results:

The goal of this study was to find a DRM Story paragraph that activated the critical lure enough to be useful for our experiments— a minimum of 30% to keep it similar to the weakest widely-used DRM list (“king”; Stadler, Roediger, & McDermott, 1995). To make this determination we relied primarily on the Recognition judgments, with Free Recall rates as a secondary measure. Given that the 7-word list we employed in Study 1b did not provide useful data for Free Recall, likely because of a floor effect, we were interested in whether the Story format would produce a different trend. As such, we did not analyze or interpret Confidence or R/K/G judgments in this study.

Free Recall Between the two samples, participants demonstrated an unexpectedly high rate of reporting “guilty” on the Free Recall memory test. In Sample 1, $M_{14\text{-Word}}=0.710$ of participants false alarmed to guilty. In Sample 2, $M_{7\text{-Word}}=0.708$ of participants false alarmed. Independent sample $t$-tests revealed these two rates were not significantly different from one another ($t(91)=0.02, p=.987$).
Recognition Judgments Similar to the Free Recall measure, participants false alarmed to “guilty” at a much higher rate than was expected. In Sample 1, M\text{14-Word}=0.826 false alarmed, while in Sample 2, M\text{7-word}=0.917 false alarmed. Although these rates varied slightly, t-test comparisons revealed they were not significantly different from one another (t(91)=1.06, p=.290).

Discussion:

Overall, the “guilty” DRM Story format was far more successful than we expected. While we expected false recognition rates around 30%, participants actually false alarmed two- to three-times as often. Further, while we made no specific predictions about Free Recall false alarms, we did not expect such high rates as we found here. Recall the purpose of Study 2a was to develop a usable Story format for “guilty” for use in Study 2b. The best story, then, would be one that would maximize the chances of measuring changes in semantic activation: Based on these results, we concluded the 7-word Story was best suited. We reasoned that in a 7-word story, a single word would contribute proportionally more to the semantic activation than would a single word in the 14-word story. Further, because the 7-word and 14-word stories were not significantly different in Free Recall and Recognition false alarms, the 7-word story does not have any memory test disadvantages. Thus, we can test our main research question with a context-rich “guilty” DRM story: Can we detect differences in automatic association between specific forms of evidence and guilt?
CHAPTER 10: STUDY 2B

Design:

Between-subjects 5-group design (DRM Story Evidence form: Control, confession, DNA, fingerprint, eyewitness).

Participants:

We recruited 1188 MTurk workers. Of these, we excluded 179 for failing to follow directions and/or technical issues. Thus, our final sample included 1009 participants (204 saw the Control list; 202 the Confession list; 201 the DNA list; 200 the Eyewitness list; and 202 the Fingerprint list). Participants were primarily female (N=713; male N=296), with an average age of 38.05 (SD=12.43). Given that our previous findings had effect sizes that were small to insignificant, we assumed any observed differences as a result of our experimental manipulation would produce small-to-medium effect sizes. Thus, a-priori power analyses calculate that this sample size, approximately 200 per cell, would be sufficient to detect a small-to-medium effect size (d=0.30) with 80% power.

English was the primary language of all but 15 participants; however these 15 participants provided free responses in the free recall sections and open-ended comments sections without spelling errors and with good grammar. Further, excluding these participants did not change the pattern of results. One hundred and thirty-one of the participants held graduate degrees; 429 college degrees; 442 high school diplomas; and 7 had not finished high school.

As in Study 1b, we performed each of our analyses for participants who had served on a jury (N=168) and participants who have had a close family member or friend convicted of a serious crime (N=247). For the results of the Recognition Test, Confidence,
and R/K/G judgments, results do not differ between subsets of participants, nor from the results of the combined sample. There were, however, differences for the Free Recall responses between subsets, which we report below.

**Materials:**

*DRM “Guilty” Story.* Based on the results of Study 2a, we developed a 7-word DRM Story with the critical lure “guilty.” This version served as our Control Story. We also created four other versions, each containing a different form of evidence (confession, DNA, eyewitness, or fingerprint). These lists were created in the same manner as Study 1b: We removed the last word (“trial”) and placed the evidence word in a novel sentence in the middle of the story paragraph (see Appendix F).

*DRM “Sleep” Story.* As a control list, we had participants read the DRM “sleep” story (Dewhurst, et al. 2007). Because we wanted to keep it approximately the same length as our “guilty” story (72 words), we used the full “sleep” story that contains fourteen sleep-related words (80 words total) rather than a story containing only the first seven words (38 words total).

For the instructions and construction of the memory measures, see Appendix L.

**Procedure:**

The procedure was identical to Study 1b with one exception: the Free Recall portion asked participants to replicate the entire paragraph they read, rather than list words.

**Hypotheses:**

*H7:* The *Confession* and *DNA* stories will generate the most false alarms (both free recall and recognition), while the *Control* stories will generate the fewest false alarms. The
AUTOMATICALLY GUILTY

Eyewitness and Fingerprint stories will generate fewer than the Confession and DNA stories, but more than the Control.

H8: The Confession and DNA stories will generate the highest confidence for the critical lure of the stories, followed by the Eyewitness and Fingerprint stories. The Control story will generate the lowest confidence for the critical lure.

H9: The Confession and DNA stories will generate the most “remember” judgments and fewest “guess” judgments for false alarms, while the Control stories will generate the fewest “remember” judgments and most “guess” judgments. The Eyewitness and Fingerprint stories will generate intermediary “remember” and “guess” judgments.

Based on our previous work, we expect these differences between groups to be small effect sizes (Cohen’s $d=0.20$)

Results:

Free Response. On average, Free Recall rates were lower than Study 2a (minimum: $M_{\text{Eyewitness}}=0.50$, 95% CI=[.43, .57]; maximum: $M_{\text{DNA}}=0.66$, 95% CI=[.59, .72]), but much higher than any of our previous work using the list format. We submitted the data to a five-level ANOVA, which yielded a significant effect of list type, $F(4, 1004)=3.12$, $p=.015$, $\eta_p^2=0.012$. This main effect is driven primarily by significant differences between the DNA Story group ($M_{\text{DNA}}=0.66$, 95% CI=[.59, .72]) and the Control Story group ($M_{\text{Control}}=0.51$, 95% CI=[.45, .58], $t(403)=2.92$, $p=.004$, $d=0.313$ 95% CI[0.254, 0.372]) and the Eyewitness Story group ($M_{\text{Eyewitness}}=0.50$, 95% CI=[.43, .57], $t(399)=3.21$, $p=.001$, $d=0.328$, 95% CI=[0.281. 0.376]. All other follow-up comparisons were not significant, $p=.72,.768$ (Appendix H)
It is worth noting that the main effect for List Type did not reach statistical significance for the participants who had served on a jury ($N=168$; $F(4, 163)=1.31$, $p=.267$), nor for the participants with a close friend or family member convicted of a felony ($N=247$; $F(4, 242)=1.47$, $p=.213$). For both of these groups, however, the results showed the same general trend, with participants who viewed the DNA story showing high rates of false recall (jury subgroup, $M=0.60$, 95% CI [.43, .77]; felony subgroup, $M=0.60$, 95% CI [.47, .74]) compared to all other list types except for the Fingerprint List (jury subgroup, $M=0.64$, 95% CI [.46, .81]; felony subgroup, $M=0.63$, 95% CI [.49, .78]). Because the trends were very similar, and the subsets represent a fraction of the sample size we predicted we would need, we do not think these differences between subsets are meaningful.

To summarize, here our hypotheses were partially supported: Participants who read the DNA Story did, in fact, show significantly more false alarms than two of the other
groups. However, participants in the *Confession* group did not false alarm more than other groups.

*Recognition Judgments.* There was very little variability in false alarming to “guilty” on the *Recognition* test between groups: minimum, $M_{\text{Control}}=0.84, 95\% \text{ CI}=[.79, .89]$; maximum: $M_{\text{DNA}}=0.89, 95\% \text{ CI}=[.84, .93]$ and $M_{\text{Eyewitness}}=0.89, 95\% \text{ CI}=[.84, .93]$. We submitted recognition judgments to a five-level ANOVA, which yielded no main effect, $F(4, 1004)=0.65, p=.630$. Unsurprisingly, follow-up $t$-test comparisons yielded no significant differences between any of the groups, $p=[.214, .986]$. Thus, the data provided no evidence for our hypotheses: neither the *DNA* nor *Confession* Stories caused more false alarms on the Recognition Test than other evidence Stories. While the *Control* Story did cause the fewest false alarms, it was not significantly lower than any other groups.

![Figure 6. Study 2b proportion of “Old” responses on Recognition Test](image)

*Confidence Judgments.* The Story Format produced higher Confidence judgments for “guilty” than we found using the List format in Study 1b (minimum: $M_{\text{Control}}=4.27, 95\%$
AUTOMATICALLY GUILTY

CI=[4.13, 4.41]: maximum: $M_{\text{Fingerprint}}=4.47$, 95% CI=[4.35, 4.59]). We submitted the data to a five-level ANOVA, which did not yield a significant main effect, $F(4, 1004)=1.36$, $p=.246$. There was, however, one significant difference in the follow-up $t$-tests: Participants in the Fingerprint Story group were more confident in their decisions about the critical lure than were participants in the Control Story group, $t(404)=2.09$, $p=.037$, $d=0.208$, 95% CI=[0.114, 0.301]. No other comparisons between groups were statistically significant, $p=[.073, .806]$.

Figure 7. Study 2b Confidence for Recognition Test responses to “guilty”

Despite the significant difference between participants who read the Fingerprint Story and participants who read the Control Story, we did not find support for our hypotheses. Although the Control story did generate the lowest confidence ratings, it was only significantly lower than participants who saw the Fingerprint list. Similarly, participants who read the DNA and Confession stories were not the most confident about their decision for “guilty,” which is what we predicted.
As in Study 1b, we tested whether different forms of evidence on the “guilty” list led to phenomenologically different false alarms to the critical lure. We did not find any differences between groups, providing no support for our hypothesis: \( \chi^2(8, N=873)=487, p=.771 \). Follow-up comparisons did not reveal any significant differences between specific list conditions (all \( p’s>.05 \)).

Interestingly, the overall trend of R/K/G responses for all lists did differ from Study 1b. Recall that in Study 1b, participants (regardless of the list they saw) responded approximately equally with “remember” and “know” responses, with slightly fewer “guess” responses. Here, however, the vast majority of participants responded with “remember” responses, followed by “know” responses, and then very few “Guess” responses (See Table 2).

Table 2.

<table>
<thead>
<tr>
<th>List Type</th>
<th>Remember</th>
<th>Know</th>
<th>Guess</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>121</td>
<td>42</td>
<td>9</td>
<td>172</td>
</tr>
<tr>
<td>Confession</td>
<td>126</td>
<td>37</td>
<td>9</td>
<td>172</td>
</tr>
<tr>
<td>DNA</td>
<td>121</td>
<td>44</td>
<td>13</td>
<td>178</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>115</td>
<td>53</td>
<td>9</td>
<td>177</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>119</td>
<td>43</td>
<td>12</td>
<td>174</td>
</tr>
</tbody>
</table>

Discussion:

As with Study 1, Study 2 was partially successful. On the one hand, we successfully developed a “guilty” DRM Story. The majority of participants falsely recalled reading “guilty” in a paragraph about the criminal justice system—both on the Free Recall and Recognition memory tests—with high confidence. On the other hand, the different evidence lists did not result in differences for any of our memory measures.
AUTOMATICALLY GUILTY

Despite no significant findings between the different lists, we are largely encouraged by the success of the story structure. It appears that the added context is quite effective in activating participants’ network of “guilty” in such a way that, perhaps, leaves it resistant to a “search to reject” strategy that typically produces a decreased DRM effect in adults.

Although the Recognition Test responses were higher than we expected, we are particularly interested in the Free Recall rates. It seems that “guilty” is a key component to participants’ understanding of a story about the legal system. Put another way, participants cannot help but activate “guilty” when the story format provides context of a person in the legal system—something we find troubling, if not unsurprising (see the Conclusion section below for a more in depth discussion of this possibility). Further, the trend of R/K/G suggests that not only are participants activating “guilty” more easily, but so too are they activating it more strongly. Nearly 69% of participants responded “remember” for the critical lure—a strikingly high rate for a false memory. Thus, the added context of a criminal trial that the story format provides, seems to activate “guilty” more easily and in a more detailed, phenomenologically-real way.

That participants reported such a high rate of clear, phenomenologically-strong false memories also suggests a reason that we found no differences between groups: a ceiling effect. Given the notably high rates of false free recall and false recognition, it may be that strongly-associated forms of evidence cannot drive false alarms rates any higher. Indeed, this was our rationale for the use of potentially-weaker 7-word lists in Study 1. One solution to such a ceiling effect may be to design a story format that has a weaker association to “guilty.” Because less work has been done with the DRM story format than the DRM list—particularly with adult populations—the exact method of how to accomplish
AUTOMATICALLY GUILTY

this would require further testing. One option may be to utilize a story that is not so clearly about a trial; perhaps a vignette that describes both criminal and non-criminal activities. Another possibility may be using the key words in a more ambiguous context. Of course, this is easier for some words than others: “judge,” “trial,” and “court” are words we use in non-criminal contexts all the time, whereas words like “lawyer” and “crime” are fairly specific to the judicial system.

Overall, we found the story format to be a very successful paradigm for eliciting false alarms to “guilty”—but less successful for our purpose of measuring associations between “guilty” and evidence. The high rates of false alarms on the Free Recall, and the clear trend in R/K/G responses, suggest this format offers a study of guilt associations that the traditional DRM list does not. Although we did not find evidence to support our hypotheses regarding the strength of association between guilt and different forms of evidence, we think the “guilty” story can be a valuable research tool for more general investigations of how people think of, and remember, guilt.

Clearly, the additional context afforded by a story format increases semantic activation of “guilt” compared to the simpler list format. Yet, this increase in context did not make differential associations between evidence and guilt easier to detect. One potential reason for the lack of differences may be that our participants did not consider any of the evidence forms to be different in the context in the trial—all evidence indicated “guilt,” regardless of the degree. Our next step, then, was to influence context in a different way by exerting more control over how people viewed the different forms of evidence. Rather than exploit the overall context of “guilty,” we can instead influence the association between
evidence and guilt more specifically—that is, by providing contextual information that influences how people think about the form of evidence via priming.
In the real world, high-quality evidence should lead to more convictions than low-quality evidence. Jury decision-making research has, to some extent, supported this premise. Several studies have found that when expert testimony is present, or fully explained, participants show increased sensitivity to the strengths and weaknesses of such evidence (e.g., Cutler, Dexter, & Penrod, 1989; Kovera, Gresham, Borgida, Gray, & Regan, 1997) Further, in a meta-analytic review of 206 studies using deliberating jurors, Devine, Clayton, Dunford, Seying and Price (2000) found strength of evidence was one of the strongest predictors of a guilty verdict: The stronger the evidence, the more likely the jury concluded the defendant was guilty. However, it is worth noting that several studies (e.g., Levett & Kovera, 2008; McAuliff & Duckworth, 2010) have found that jurors are sometimes unable to differentiate between good and bad evidence.

Taken together, these results suggest that any method that would cause someone to view a specific type of evidence as higher quality would cause a corresponding increase in the association between that evidence and guilt. Conversely, any manipulation that leads someone to see a particular form of evidence as being of poorer quality should equally decrease the association between that evidence and guilt. Thus, if we can influence how people view the quality of a specific form of evidence, such strengthening or weakening of the association between that evidence and guilt could be measurable with our DRM paradigm in the form of more (or less) false recall. Priming is one such method that may cause such a shift in association strength, because it has been used successfully to influence how people false alarm on a DRM task (Lenton et al., 2001; Takarangi et al., 2007). In our
study, the primes will be information regarding the scientific validity of each form of evidence.

Scientists have established standards for forensic evidence: some versions are of high validity, such as DNA; others are characterized as unscientific, such as bitemark evidence (National Academy of Sciences, 2009; 2014). However, previous research has demonstrated contextual information about forensic evidence can alter how people regard that evidence. Lawson (2014) found that participants who read illustrative stories about legal cases that included facts about evidence altered their views about the validity of evidence. Lawson was primarily motivated in how pretrial publicity about evidence quality can influence a case. Whereas pretrial publicity studies have normally focused on guilt judgments and more direct judgments of the defendant (e.g., Ruva & McEvoy, 2008), pretrial information can also affect the case indirectly by influencing opinions of evidence. Here, we seek to accomplish something similar by directly influencing how our participants view evidence in an attempt to alter that evidence’s connection to guilt.

It is likely, however, that some forms of evidence are more resistant to priming than others. Forms of evidence that are more closely associated with guilty, such as confession and DNA, would likely require extremely strong manipulations to influence that association. For example, as reviewed in Chapter 1, both real cases and laboratory research have demonstrated a confession very likely leads to a guilty verdict (Kassin & Sukel, 1997) and is more likely to lead to a guilty verdict than other forms of evidence (Kassin and Neumann, 1997). Weaker-associated evidence-types, such as eyewitness and fingerprints, that sometimes (but do not always) lead to a guilty verdict, may be more susceptible to such priming, because a person’s belief in the reliability of these forms of evidence may be
easier to shift. Therefore, in Study 3 we ask, can priming participants with evidence quality affect the evidence's association with guilt in a DRM paradigm?
CHAPTER 12: STUDY 3A

Design:

Because the goal of this Study was to generate effective negative- and positive-evidence primes, there was no experimental manipulation. Instead, participants saw either two positive- and two negative-evidence lists (Sample 1) or one of the eight possible primes (Samples 2 & 3). We varied the number of primes participants saw after Sample 1 due to concerns over anchoring-and-adjustment (Tversky & Kahneman, 1974) for particularly strong and weak forms of evidence (see Results for the full rationale).

Participants:

In Sample 1, we collected data from 75 MTurk workers. Of these, we excluded 6 for failing to follow directions and/or technical issues. Thus, our final sample included 69 participants (35 saw the DNA-, Confession+, Fingerprint+, and Eyewitness- primes; 34 saw the DNA+, Confession-, Fingerprint-, and Eyewitness- primes). Participants were 27 female, 25 male, with an average age of 37.99 (SD=12.02). English was the primary language of all participants. 11 held graduate degrees, 30 college degrees, and 27 high school diplomas, and 1 did not complete high school.³

In Sample 2, we collected data from 81 participants and excluded 14. Our final sample included 67 participants (approximately 8 participants saw each of the eight primes; see Table 4 for exact N per cell), 39 of which were female and 28 male, with an average age of 38.40 (SD=11.73). English was the primary language of all participants; 8

³ This is the same sample as Study 2a Sample 1. After reading the Story Format and taking the memory test, these participants then read the evidence primes for this study.
AUTOMATICALLY GUILTY

held graduate degrees, 27 college degrees, 29 high school diplomas, and 3 did not complete high school.

Sample 3 was composed of 163 participants, 8 of whom we excluded, leaving 155 participants for analyses (approximately 19 participants per prime; see Table 5 for exact N per cell). 106 of these participants were female and 49 were male; on average, they were 37.72 years old (SD=11.71). English was the primary language of all participants. 20 participants held graduate degrees; 64 held college degrees; 68 had a high school diploma; and 3 did not finish high school.

Materials:

Evidence Ratings. Measures In order to assess the efficacy of the primes (that is, to determine whether a negative prime decreased participants’ trust in that specific evidence, and a positive prime increased participants’ trust), we asked participants to rate each form of evidence in three ways: 1) how reliable (how often it is correct or wrong; 10=never wrong, 0=always wrong); 2) how convincing (how useful it would be for the participant in making a guilt determination if they were on a jury; 10=completely convincing, 0=not convincing at all) they believed it was. Participants completed this rating twice: once before reading the prime, and then again after reading the prime; 3) immediately after reading the prime, participants responded to the question: “How did the paragraph you just read affect your opinion of [the form of evidence]?” on a scale of -2 to +2 (-2=I trust it a lot less; -1=I trust it a little less; 0=no change, I trust it the same; 1=I trust it a little more; 2=I trust it a lot more).

Evidence Primes. We developed eight short vignettes that framed the four pieces of evidence we use in our DRM lists and stories (positive and negative versions for confession,
DNA, eyewitness, and fingerprint). The positive vignettes explained how the key piece of evidence led to a perpetrator of a murder being correctly convicted, while the negative vignettes explained how the key piece of evidence led to an innocent person being wrongfully convicted of the murder. The vignettes for each evidence form differed in some aspects, such as the location of the murder, names of the perpetrator, and how the police conducted the preliminary investigation. Across the three samples, we slightly varied the content of the vignettes to achieve the desired shifts in reliability, credibility, and opinions in the evidence. For example (changes underlined):

**Sample 1**

Tom Watkins, of Milwaukee, WI, was recently convicted of first degree murder after a two-day trial. Police identified Watkins as a suspect in the murder of another Milwaukee man, Sam Perkins. After Perkins did not report for work for two days, police went to his house where they found his body. Police arrested Watkins after a preliminary investigation revealed a connection between Watkins and Perkins.

At trial, Watkins was convicted based primarily on his confession to the crime. The jury found Watkins’ confession to be very compelling evidence, because it contained many accurate details, and the police gathered it using correct, standard interrogation methods, making it very reliable and likely true. Further, the jury heard expert testimony that a person will rarely confess to a crime they did not commit. Therefore, Watkins’ confession was key in making sure the right person was brought to justice.

**Sample 2**

Tom Watkins, of Milwaukee, WI, was recently convicted of first degree murder after a two-day trial. Police identified Watkins as a suspect in the murder of another Milwaukee man, Sam Perkins. After Perkins did not report for work for two days, police went to his house where they found his body. Police arrested Watkins after a preliminary investigation revealed a connection between Watkins and Perkins.

At trial, Watkins was convicted based primarily on his confession to the crime. The jury found Watkins’ confession to be very compelling evidence, because it contained many accurate details, and the police gathered it using correct, standard interrogation methods, making it very reliable and likely true. Further, the jury heard expert testimony that a person will rarely confess to a crime they did not commit. In fact, confessions help to catch many perpetrators. Therefore, Watkins’ confession was key in making sure the right person was brought to justice.
Sample 3

Tom Watkins, of Milwaukee, WI, was recently convicted of first degree murder after a two-day trial. Police identified Watkins as a suspect in the murder of another Milwaukee man, Sam Perkins. After Perkins did not report for work for two days, police went to his house where they found his body. Police arrested Watkins after a preliminary investigation revealed a connection between Watkins and Perkins.

At trial, Watkins was convicted based primarily on his confession to the crime. The jury found Watkins’ confession to be very compelling evidence, because it contained many accurate details, and the police gathered it using correct, standard interrogation methods, making it very reliable and likely true. Further, the jury heard expert testimony that a person will rarely confess to a crime they did not commit. In fact, confessions help to catch many perpetrators. After the trial, the judge offered an opinion that a confession is reliable and is often correct, as it was in this case. The judge also wrote that confession evidence should be very convincing that the defendant committed the crime. Therefore, Watkins’ confession was key in making sure the right person was brought to justice.

For a full list of the vignettes, see Appendix I.

Procedure:

Sample 1 participants rated all four forms of evidence on their reliability and convincingness. Then, they read four of the eight possible primes (two positive and two negative, one for each form of evidence, randomized order) and rated how their opinion of the evidence changed immediately after each prime. After reading the four primes, participants again rated the reliability and convincingness of all four forms of evidence a second time before being debriefed.

Sample 2 and 3 participants rated only one form of evidence determined by random assignment. After rating the evidence’s reliability and convincingness, participants read one prime, rated how it affected their opinion of the evidence, and then again rated the evidence’s reliability and convincingness before being debriefed.
AUTOMATICALLY GUILTY

Results:

To determine whether our primes were successful, we focused on three measurements: the change in reliability ratings between pre- and post-prime ratings, the change in convincingness ratings between pre- and post-prime ratings, and the post-prime ratings of how the primes changed participants' opinion of the evidence.

Sample 1. Our first version of the evidence primes was not successful, but showed promise. On some measures, for some of the evidence forms, participants demonstrated significant differences between the positive and negative primes (e.g., change in opinion of DNA, confession reliability and convincingness); however in some cases, the differences were not significantly different between positive and negative primes (eyewitness reliability and credibility) or the differences were negatively skewed. For example, despite significant differences between groups, participants rated DNA reliability, DNA convincingness, fingerprint reliability, and fingerprint convincingness more negatively after viewing the positive prime (Table 3).
Table 3

*Study 3a Sample 1 Priming Pilot Evidence Ratings*

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Priming Quality (N)</th>
<th>Trust in Evidence</th>
<th>ΔReliability</th>
<th>ΔConvincingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confession</td>
<td>Positive (34)</td>
<td>0.23 (0.73)</td>
<td>0.20 (1.57)</td>
<td>0.06 (1.43)</td>
</tr>
<tr>
<td></td>
<td>Negative (35)</td>
<td>-0.62 (0.88)</td>
<td>-0.88 (1.59)</td>
<td>-0.79 (1.61)</td>
</tr>
<tr>
<td>DNA</td>
<td>Positive (34)</td>
<td>0.12 (0.92)</td>
<td>-0.44 (1.16)</td>
<td>-0.21 (1.10)</td>
</tr>
<tr>
<td></td>
<td>Negative (35)</td>
<td>-0.51 (0.84)</td>
<td>-1.63 (2.66)</td>
<td>-1.37 (2.30)</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Positive (35)</td>
<td>0.15 (0.82)</td>
<td>0.32 (1.41)</td>
<td>0.15 (1.74)</td>
</tr>
<tr>
<td></td>
<td>Negative (34)</td>
<td>-0.66 (0.87)</td>
<td>-0.26 (1.36)</td>
<td>-0.71 (1.98)</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Positive (35)</td>
<td>0.37 (0.73)</td>
<td>-2.00 (2.44)</td>
<td>-0.51 (1.42)</td>
</tr>
<tr>
<td></td>
<td>Negative (34)</td>
<td>-0.79 (0.77)</td>
<td>-3.26 (2.70)</td>
<td>-1.50 (1.54)</td>
</tr>
</tbody>
</table>

*Note.* Bolded values are significantly different (p<.05) between the Positive and Negative groups.

Our method, however, presented a potential confound for these variables. Because participants viewed four primes and rating all four forms of evidence, we suspected that it was possible the primes were interacting in such a way as to decrease perceived credibility of all forms of evidence, not just those in the negative primes. For example, participants may have read the negative DNA prime and thought if DNA can be flawed, *all* evidence can be flawed. Therefore, we decided to collect another sample with two alterations. First, we tweaked the wording of our primes (see Appendix I). Second, we only showed participants one prime and had them rate only the relevant form of evidence; thus, they were not (with or without awareness) comparing one form of evidence to another.

*Sample 2.* Overall, the adjustments to the procedure (only one prime) and tweaks to the wording of the primes moved us closer to effective primes. As seen below (Table 4), all primes resulted in rating changes in the correct direction. Unfortunately, many of the ratings were not significantly different between participants in the positive prime and participants in the negative prime groups. Additionally, some of the primes yielded no
change in ratings (e.g., positive prime confession reliability, and positive prime DNA reliability). One possible reason for such small, sometimes non-existent, differences for the *credibility* and *convincingness* scores was that the pre-prime rating of these aspects was anchoring the post-prime ratings. As such, we only measured post-prime ratings in Sample 3, and again tweaked the wording of our primes in the hope of strengthening the effect.

Table 4

*Study 3a Sample 2 Priming Pilot Evidence Ratings*

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Priming Quality (N)</th>
<th>Trust in Evidence</th>
<th>ΔReliability</th>
<th>ΔConvincingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confession</td>
<td>Positive (7)</td>
<td><strong>0.57 (0.54)</strong></td>
<td>0.00 (1.41)</td>
<td>-0.14 (1.35)</td>
</tr>
<tr>
<td></td>
<td>Negative (9)</td>
<td><strong>-0.78 (-0.78)</strong></td>
<td>-1.11 (1.27)</td>
<td>-0.78 (0.97)</td>
</tr>
<tr>
<td>DNA</td>
<td>Positive (9)</td>
<td><strong>0.00 (0.50)</strong></td>
<td>0.00 (0.00)</td>
<td><strong>-0.11 (0.78)</strong></td>
</tr>
<tr>
<td></td>
<td>Negative (7)</td>
<td><strong>-1.00 (0.58)</strong></td>
<td>-3.29 (2.36)</td>
<td><strong>-3.00 (2.24)</strong></td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Positive (7)</td>
<td>-0.13 (0.99)</td>
<td><strong>0.13 (1.55)</strong></td>
<td>0.25 (0.46)</td>
</tr>
<tr>
<td></td>
<td>Negative (8)</td>
<td>-0.73 (1.01)</td>
<td><strong>-1.45 (2.84)</strong></td>
<td>-2.18 (3.16)</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Positive (8)</td>
<td><strong>0.38 (0.92)</strong></td>
<td>0.00 (0.93)</td>
<td>0.25 (1.67)</td>
</tr>
<tr>
<td></td>
<td>Negative (8)</td>
<td><strong>-0.75 (0.71)</strong></td>
<td>-1.25 (1.75)</td>
<td>-1.38 (2.26)</td>
</tr>
</tbody>
</table>

*Note.* Bolded values are significantly different (*p*<.05) between the Positive and Negative groups.

*Sample 3.* As seen below (Table 5), the third iteration of our priming paragraphs were successful in creating differences between the positive- and negative-prime for each form of evidence. Because we no longer measured pre- and post-prime ratings, the goal for *credibility* and *convincingness* ratings was that the positive-prime participants rated the evidence higher than the negative-prime participants: indeed, for every form of evidence we found this trend. Further, negative-prime participants remarked that they trusted the evidence less, whereas positive-prime participants expressed trusting the evidence more.
Table 5

*Study 3a Sample 3 Priming Pilot Evidence Ratings*

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Priming Type (N)</th>
<th>Trust in Evidence</th>
<th>Reliability</th>
<th>Convincingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confession</td>
<td>Positive (21)</td>
<td>0.38 (0.87)</td>
<td>7.24 (1.81)</td>
<td>7.29 (2.31)</td>
</tr>
<tr>
<td></td>
<td>Negative (14)</td>
<td>-0.93 (0.83)</td>
<td>4.00 (2.11)</td>
<td>5.79 (2.87)</td>
</tr>
<tr>
<td>DNA</td>
<td>Positive (18)</td>
<td>0.22 (0.94)</td>
<td>7.94 (1.63)</td>
<td>7.72 (1.67)</td>
</tr>
<tr>
<td></td>
<td>Negative (21)</td>
<td>-0.76 (0.89)</td>
<td>6.81 (1.83)</td>
<td>6.62 (1.75)</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Positive (18)</td>
<td>-0.22 (0.65)</td>
<td>6.61 (2.17)</td>
<td>6.50 (2.33)</td>
</tr>
<tr>
<td></td>
<td>Negative (21)</td>
<td>-1.14 (0.79)</td>
<td>3.00 (2.74)</td>
<td>3.10 (2.51)</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Positive (20)</td>
<td>0.35 (1.04)</td>
<td>8.15 (1.69)</td>
<td>7.30 (2.08)</td>
</tr>
<tr>
<td></td>
<td>Negative (22)</td>
<td>-1.05 (0.76)</td>
<td>4.86 (2.36)</td>
<td>4.45 (2.18)</td>
</tr>
</tbody>
</table>

*Note.* Bolded values are significantly different (*p*<.05) between the Positive and Negative groups.

**Discussion:**

It is worth mentioning that in a few cases, the primes were not successful according to strict statistical convention. For instance, the positive- and negative-primes did not result in significant differences for confession *convincingness* (*p*=.096), and DNA convincingness (*p*=.052). This finding is not completely surprising: Given that DNA and confessions are seen as strong indicators of guilt, it logically follows that a single vignette may do little to shift participants’ opinion of the evidence’s credibility as a whole. Similarly, the eyewitness positive prime produced slightly negative change for participants’ opinion of the evidence after reading the vignette—even though it was significantly higher than positive participants’ ratings. But again, this result does not necessarily invalidate the usefulness of the prime: it is a very small negative rating (*M*=-0.22 on a 2 to -2 scale).

Further, because eyewitnesses lead to so many wrongful convictions, our participants may
be inherently skeptical of any court case in which an eyewitness is the only piece of evidence, making it unlikely they will form more positive views of eyewitness evidence.

These minor problems aside, we believed that Sample 3 provided sufficient evidence to warrant use of these primes in Study 3b, in which we measured how priming evidence quality information influences responses to our evidence “guilty” DRM lists.
CHAPTER 13: STUDY 3B

Design:

Between-subjects 3(Prime Type: positive, negative, neutral) x 4(Evidence DRM list: confession, DNA, fingerprint, eyewitness) + 1 (no-prime control list).

Participants:

We recruited 804 participants for Study 3b. Of these, 150 were excluded for failure to follow directions and/or technical issues, leaving 654 participants for analyses (approximately 50 participants per cell). An a-priori power analysis calculated this sample size was sufficient to detect a small main effect ($f=0.15$) with 80% power. Participants had an average age of 38.07 years (SD=13.09); 426 were female and 228 were male. 94 held a graduate degree, 285 a college degree, 274 a high school diploma, and 1 did not complete high school. English was the primary language of all but 5 participants; however, these 5 participants responded to the free recall and free response sections of the study with good grammar and without any errors.

Of our final 654 participants, 127 reported having served on a jury, and 146 reported having a close friend or family member convicted of a felony. Separate analyses were performed for these subgroups, and they did not vary from each other, their counterpart subgroups, or the overall sample for any of our analyses. Thus, we report the combined data.

Materials:

*Positive and Negative Evidence Primes.* Participants (except Control group) read one of eight possible primes (positive or negative for confession, DNA, eyewitness, or fingerprint) developed in Study 3a (See Appendix I).
**AUTOMATICALLY GUILTY**

*Filler Vignettes.* In order to disguise the link between the prime and subsequent DRM list, the experimentally-relevant evidence prime was embedded between two unrelated, but stylistically similar, filler primes (See Appendix I).

*“Guilty” DRM List.* Each participant read one of five possible 7-word “Guilty” DRM lists— a Control List, Confession List, DNA List, Eyewitness List, or Fingerprint List. These five lists were the same 7-word list developed and used in Study 1 (Appendix C).

For the instructions and construction of the memory measures, see Appendix L.

**Procedure:**

Participants agreed to participate in what they told was an MTurk HIT composed of two separate studies. In fact, the two studies were related: Part I was the Priming Phase, and Part 2 was the DRM Phase. This method of ostensibly running two unrelated studies has proven successful in the past for our lab (Lawson, 2014) at disguising the purpose of primes from participants. After providing informed consent, participants were randomly assigned to one of the 13 possible experimental cells, which dictated which priming paragraph and which evidence “guilty” DRM list they saw.

In the Priming Phase, participants read an evidence priming paragraph (determined by their group assignment), preceded and followed by similar, but unrelated, filler paragraphs to maintain the cover story. Participants in the Control group and neutral prime conditions read only the filler paragraphs.

After completing the Priming Phase, participants read a brief explanation of “the second study” to maintain the cover story. After reading the instructions, participants completed the DRM Phase, which was identical to the “guilty” list portion of Study 1b. Participants viewed one of the five possible lists that corresponded to the evidence in the
AUTOMATICALLY GUILTY

priming vignette from the Priming Phase (Control participants viewed the control list). For example, participants who were in the Positive Confession and Negative Confession conditions read different primes, but both viewed the “guilty” DRM list containing “confession.” After viewing the 7-word “guilty” list, participants completed a five-minute filler task, followed by a Free Recall memory test and a Recognition Test in which participants rated words as Old or New, rated their confidence in their decision, and made a “remember”, “know”, or “guess” judgment if they judged a word as “Old.” After completing the memory tests, participants answered the exclusionary questions, including what they thought the study was about, and then were debriefed.

Hypotheses:

\( H_{10} \): There will be a main effect for Prime Type, such that participants in the positive prime condition will show more false alarms (both free and recognition) for “guilty,” higher confidence for “guilty,” and more “remember” judgments than the participants who in the no prime condition, which in turn will show more false alarms, higher confidence, and more “remember” judgments than those in the negative prime condition.

\( H_{11} \): There will also be a main effect for Evidence DRM List, such that participants who read the Confession or DNA list will show more false alarms (both free and recognition) for “guilty,” higher confidence for “guilty,” and more “remember” judgments than the participants who read the Eyewitness or Fingerprint list, which in turn will show more false alarms, higher confidence, and more “remember” judgments than those in the Control list condition.

\( H_{12} \): There will be a significant Prime Type x Evidence DRM List interaction, such that participants who read the Eyewitness or Fingerprint list will be more affected by the
AUTOMATICALLY GUILTY

positive and negative primes than participants who read the *Confession* or *DNA* lists. That is, participants who read the *Eyewitness* or *Fingerprint* list will show greater (or fewer) false alarms, greater (or less) confidence for “guilty,” and more (or fewer) “remember” judgments compared to those in the corresponding neutral prime conditions than those who read the *Confession* or *DNA* lists.

Based on our previous work, we expect these differences between groups to be small effect sizes (Cohen’s $d=0.20$). Further, we expect small main effects as well (Cohen’s $f=0.25$)

Results:

For Free Recall, Recognition Test, and Confidence Judgment results, data were submitted to a 3(prime type: positive, negative, neutral) x 4 (evidence type: confession, DNA, eyewitness, fingerprint) ANOVA. These analyses tested for the hypothesized Priming main effect, Evidence List main effect, and Priming*Evidence List interaction. Because there was no Control list equivalent to a positive or negative prime (focusing on the credibility aspect of a single word on the list), we could not fullyfactorially-cross the Control condition, and thus excluded it from analyses.

*Free Response* Overall, Free Response rates of “guilty” were fairly low: maximum $M_{Control}=0.18$, 95% CI=[.07, .028], $M_{EyewitnessNegative}=0.18$, 95% CI=[.07, .29], and $M_{FingerprintNegative}=0.18$, 95% CI=[.07, .29]: minimum $M_{DNA}=0.04$, 95% CI=[0.00, 0.10]. These free recall rates are much lower than those found in Study 2b, but similar to those found in Study 1b—a trend that is not surprising, given that we used the 7-word list format rather than the story format.
We found no support for a main effect of Prime Type, $F(2, 591)=1.79$, $p=.169$; no main effect of Evidence List, $F(3, 591)=0.36$, $p=.784$; and no interaction between Prime Type and Evidence List, $F(6, 591)=1.18$, $p=.313$.

**Figure 8.** Study 3b Proportion of “guilty” responses to Free Recall Test

In order to investigate the possibility that priming may have affected some forms of evidence more than others, we conducted a number of follow-up t-tests. Specifically, we compared the free recall of “guilty” between the positive and negative conditions for each evidence list. This resulted in four t-tests: *Positive* and *Negative Confession*; *Positive* and *Negative DNA*; *Positive* and *Negative Eyewitness*; and *Positive* and Negative *Fingerprint*. We reasoned that if priming had any effect on the activation of guilty, these comparisons would be the most likely place to detect a significant difference. However, all of these t-tests yielded non-significant results ($p=[.130, .752]$; See Appendix K). In fact, the highest rates of guilty free recall occurred in the *Negative Eyewitness* and *Negative Fingerprint* conditions. Thus, we found no support for any of our hypotheses in the Free Recall responses.
Recognition Judgments  Similar to the Free Recall rates, the Old/New judgments for this study were similar to those in Study 1b; maximum: $M_{Positive\text{Confession}}=0.74$, 95% CI=[.61, .87]; minimum: $M_{Negative\text{Eyewitness}}=0.52$, 95% CI=[.38, .66] (Figure 9).

Similar to Free Recall results, we found no support for our hypotheses. There was no main effect of Prime Type, $F(2, 591)=1.13$, $p=.325$, no main effect of Evidence List, $F(3, 591)=0.74$, $p=.528$, nor a significant interaction between Prime Type and Evidence List, $F(6, 591)=0.14$, $p=.737$. We again conducted comparison $t$-tests between the Positive and Negative conditions for each form of evidence. None of these tests yielded significant differences, $p=[.060, .742]$. It does bear mentioning that these data trend in the predicted direction, however. Participants in the positive priming conditions false alarmed to guilty more than participants in the negative priming conditions for the Confession lists ($M_{Positive\text{Confession}}=0.74$, 95% CI=[.61, .87], $M_{Negative\text{Confession}}=0.56$, 95% CI=[.42, .70]; $t(98)=1.90$, $p=.060$), DNA lists ($M_{Positive\text{DNA}}=0.62$, 95% CI=[.48, .76], $M_{Negative\text{DNA}}=0.56$, 95% CI=[.42, .70]; $t(98)=0.61$, $p=.547$), and Eyewitness lists ($M_{Positive\text{Eyewitness}}=0.60$, 95% CI=[.46, .74], $M_{Negative\text{Eyewitness}}=0.52$, 95% CI=[.38, .66]; $t(98)=0.80$, $p=.425$). Participants who saw the Fingerprint list showed the opposite pattern, ($M_{Positive\text{Fingerprint}}=0.61$, 95% CI=[.47, .75], $M_{Negative\text{Fingerprint}}=0.64$, 95% CI=[.50, .78]; $t(98)=0.33$, $p=.742$). Although, again, these differences are not significant, they suggest that priming may more reliably influence Old/New judgments than Free Recall judgments.
Confidence Judgments. Confidence judgments were similar to those in Study 1b, ranging from a maximum: $M_{\text{Control}}=3.98$, 95% CI=[3.67, 4.29], to minimum: $M_{\text{NegativeFingerprint}}=3.33$, 95% CI=[2.96, 3.71]. We found no significant main effect for Prime Type, $F(2, 591)=0.67$, $p=.499$, no main effect for Evidence List, $F(3, 591)=0.40$, $p=.754$, and no significant Prime Type * Evidence List interaction, $F(6, 591)=1.51$, $p=.173$. Follow-up $t$-tests between the Positive and Negative conditions for each form of evidence also yielded no significant differences: $p=[.122, .719]$ (Appendix K). Thus, we found no support for any of our hypotheses.

Interestingly, participants in the positive prime conditions expressed the lowest confidence compared to their neutral and negative counterparts for the Confession lists, Eyewitness lists, and Fingerprint lists. Participants in the positive DNA condition, on the other hand, expressed higher confidence than those in the neutral DNA or negative DNA conditions. Although none of these differences are statistically significant, it a surprising
trend that the priming conditions we expected to yield the highest confidence in recognition for “guilty” in fact yield the lowest (Figure 10).

Figure 10. Study 3b confidence for Recognition Test responses to “guilty”

R/K/G Results. Due to the ordinal nature of the remember/know/guess responses and the two multi-level independent variables, the chi-square test used for Studies 1B and 2B was not an appropriate statistical test to use here. Specifically, we are interested in the possibility of the interaction between Prime Type and Evidence List. Therefore, we performed an ordinal regression with our independent variables, Prime Type and Evidence List, as our predictors. This ordinal regression yielded a poor model fit for the data, $\chi^2(11)=7.19, p=.784$. Indeed, the pseudo R-square revealed the model explained very little of the variance (Nagelkerke $R^2 = .022$), while the Test of Parallel Lines provided no evidence to reject the null hypothesis that the location parameters differed as a function of experimental assignment ($\chi^2(11)=10.16, p=.516$). Therefore, we can conclude that neither the Prime Type nor the Evidence List affected the phenomenological strength of false
recognition of “guilty.” Further, there was no evidence of any interaction of the two independent variables (Table 6).

Table 6

Study 3b R/K/G Ratings for “Old” guilty ratings

<table>
<thead>
<tr>
<th>List Type</th>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>K</td>
<td>G</td>
</tr>
<tr>
<td>Control</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Confession</td>
<td>13</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>DNA</td>
<td>14</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>9</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>10</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

Discussion:

Despite previous research showing strong effects of priming influencing how people respond to DRM lists (Lenton, et al., 2001; Takarangi et al., 2007), we found no support that our evidence quality primes affected the activation of the “guilty” semantic network. Participants did not show shifts in falsely recalling or recognizing “guilty,” nor did they show any changes in confidence or phenomenological memory quality as a result of being positively or negatively primed. Further, all four of our evidence forms showed no susceptibility to priming.

There are several explanations for why we saw no effects here. One possibility is that our primes, although successful in influencing explicit ratings of evidence quality, may have lost their potency by the time participants viewed the DRM list. In pilot testing, participants rated the evidence immediately after viewing the prime. However in Study 3b, participants read a filler prime and DRM instructions before viewing the list, resulting in approximately three minutes elapsing between the prime and reading the list, and over
eight minutes between the prime and the free recall test. It may be that the admittedly
small effects we measured in pilot testing were insufficient to influence semantic activation
while viewing the list; it is also possible that the effects of the primes we did observe in 3a
were too short-lived and were thus over by the time participants viewed the list.

Another possibility is that changing the relationship between one word (the
evidence word) and guilty is not a strong enough effect to influence overall semantic
activation. Indeed, past research has primed the entire theme of the list; Lenton et al.
(2001) primed gender expectations, and Takarangi et al., (2007) primed violence. Here, we
focused our prime on a single word. If our speculation accounting for the lack of list
differences in Studies 1 and 2 is true that changing one word has a small, even negligible
effect on semantic activation, it also stands to reason that strengthening or weakening a
single word will also have no effect.

Our data did show several trends worth noting. First, negative and positive primes
did show the predicted trend for Recognition Test responses. Specifically, participants in
the Positive Confession condition false alarmed to “guilty” far more often than participants
in the Negative Confession condition. Although the t-test was not significant, it is possible
this result would be significant with more participants. Here we had only 50 participants
per group, whereas in past research we use 200 participants per cell. Given the exploratory
nature of this experiment, future research could use these results as a guide for more
specific, greater-powered use of this paradigm.

Despite the lack of significant findings here, we believe this paradigm is ripe for
further research. For example, our research question here was whether priming affected
the semantic relationship between a specific form of evidence and guilt. What we do not
know is how priming may affect the “guilty” list in general. For example, would more thematic (as opposed to single word) priming influence activation of guilty? Perhaps instead of focusing on a single form of evidence, a priming paragraph could depict a more detailed right (positive) or wrongful (negative) conviction. If more globally-thematic priming influences activation of guilty, how would such priming compare to priming other, more traditional DRM lists?
CHAPTER 14: GENERAL DISCUSSION

Because evidence sometimes leads to a conviction despite overwhelming proof of innocence, we suspect that certain forms of evidence are strongly associated with guilt. As such, hearing strongly guilt-associated evidence automatically and strongly activates the thought of guilt—and later, people may remember “guilty” rather than the actual evidence. Here, we tested for such automatic association with the DRM paradigm (Roediger & McDermott, 1995) in a novel way. Specifically, we asked participants to read a group of words related to “guilty” (but not “guilty” itself). Then, on a memory test for the words they saw, participants incorrectly report seeing “guilty” because of a semantic activation error: all of the “guilty”-related words activate the semantic network containing “guilty” and participants mistake thinking of “guilty” for actually having read it. The stronger the connection between the words participant read and “guilty,” the more likely they are to falsely recall reading “guilty.” Therefore, the rate of false recall is an indirect measure of the strength of automatic association—a measure we can use to detect the strength of automatic association between evidence and guilt.

In this set of studies, we had participants read our “guilty” list, but varied a single word between participants: namely, one form of evidence. Activation monitoring theory suggests that participants who read the group of words containing evidence strongly associated with “guilty” would falsely report having read “guilty” more than participants reading more weakly associated evidence. For instance, if our hypothesis was correct, participants reading “confession” (which often results in a conviction, even when it should not; Kassin & Neumann, 1997; Kassin & Sukel, 1997) or “DNA” (the “gold standard” of evidence; National Academy of Sciences, 2009) should falsely recall “guilty” very often.
Conversely, participants who read “eyewitness” (a form of evidence that is demonstrably problematic; National Academy of Sciences, 2014) or “fingerprint” (which follows a general methodology, but not always reliable; National Academy of Sciences, 2009) should have falsely recalled “guilty” less often. In Study 1, we tested our hypothesis by listing six of our “guilty” words and one form of evidence (the 7-word list format). In Study 2, we embedded the six “guilty” words and one evidence word in a paragraph about the criminal justice system (the “DRM Story” format) to increase the contextual information and potentially alter activation of “guilty.” Finally, in Study 3, we influenced context in a different way by providing participants with a vignette that described a piece of evidence as either reliable, good evidence or unreliable, flawed evidence. We used this positive or negative information to make that form of evidence more or less associated with “guilty,” —a shift which, we expected, would be detectable using the DRM paradigm.

This is, to our knowledge, the first attempt at using the DRM as a tool to specifically measure automatic associations. Although researchers have used the DRM variations we used here (a 7-word list, Gallo, 2010; the DRM Story format, Dewhurst, Pursglove, & Lewis, 2007; and using priming to alter the association network, Lenton, et al., 2001; Takarangi et al., 2007), we had to adapt each approach for a “guilty” DRM list. Each of our new DRM iterations successfully caused participants to falsely recall seeing “guilty” at rates similar to established lists that exhibit very strong DRM effects (e.g. “sleep”). However, in each of our studies, we detected no differences in false free recall, false recognition alarms, confidence, or remember/know/guess judgments between participants who saw different evidence lists. The single exception to this was the free recall rates for the “DNA” Story from Study 2b, which produced significantly higher false recall rates for “guilty” than the Control and
AUTOMATICALLY GUILTY

Eyewitness Stories. Despite this singular success, simply put, we found little evidence to support our hypotheses that different forms of evidence (confession and DNA) would yield higher false recall for “guilty” in our applications of the DRM paradigm.

Why did we fail to detect changes in response in the DRM paradigm? One possibility is that the associations between different form of evidence and guilt are not as automatic as we suspected, and instead require additional cognition in order to create the association. However, if this were the case, we would expect lists that contain a form of evidence would generate lower false recall rates than the control list, due to weakening the semantic activation with a non-associated word. Secondly, we would expect that participants who view information cuing the context of the criminal justice system (e.g., the “guilty” story, and the positive and negative evidence primes) would activate “guilty” and thus show higher false recall for “guilty” than participants who did not receive such information. In both cases, we did not find such a trend. Participants regularly responded with high rates of false alarms to the “guilty” DRM list and story. Further, participants receiving additional contextual information did not respond with more false recall than those that did not. Notably, participants who viewed the “guilty” DRM story had far more false memories for “guilty” in the free recall portion than did participants who viewed the list format; however, the rates of recognition false alarms were extremely similar. If there were no association, participants in the different Neutral priming conditions should show fewer false memories and false alarms than those in the Negative priming condition—a trend we did not find. Thus, we are confident the association is automatic, but the paradigms we used here were insufficient to measure it. What remains unclear, however, is whether different forms of evidence form automatic associations that differ in strength with “guilty.”
AUTOMATICALLY GUILTY

There are a number of potential explanations for why we failed to measure differential association strengths between different evidence and guilt. One explanation is that the one-word alteration we use here is insufficient to change the activation strength of “guilty.” Simply put, the network is activated by the first few words on the list immediately, and remaining words do little to shift the activation strength. As such, the evidence embedded in the middle of the list or story has a negligible effect on the overall activation strength. We find this explanation to be the most probable for two reasons. First, it seems that the effect of evidence type on activation strength is extremely weak, rather than non-existent. Throughout our studies we found results that trend in the hypothesized distribution of false recall, despite non-significant effects. This trend suggests that the differential association strengths with “guilty” are ever-so-slightly pushing false recall rates in the expected direction (up for confession and DNA; down for fingerprint and eyewitness). We are not saying, of course, that our hypotheses are true despite lack of statistical significance. Indeed, that each study finds a hypothesized trend in a different measurement, but these do not replicate across studies does little to paint a clear picture.

For example, participants who saw the Confession List in Study 1b and the “positive confession” prime and list in Study 3b reported more false alarms to “guilty” on the Recognition Test than their counterparts; however, in Study 2b, Confession Story participants produced among the lowest recognition false alarms. In Study 2b, the DNA Story produced the most false memories on the Free Recall test, significantly more than the Confession Story and Fingerprint Story but this finding did not replicate across any other studies or memory measures. Thus, there seems to be an effect of evidence type on
AUTOMATICALLY GUILTY

semantic activation—but it is of small size and large variability, and therefore extremely
difficult to pinpoint.

The trend in priming results from Study 3b provides a second potential reason for
why a single evidence word is insufficient to shift activation strength. Although we found
no main effect for priming, the Recognition Test results do show a general trend towards
more false alarms for the positive primes than the negative primes, albeit not at a
significant level. These results suggest that our use of the DRM paradigm is capable of
detecting changes in association strength. Particularly, it seems that the DNA list may have
been more resistant to priming than other forms of evidence—a trend we predicted
because the association with guilt is very difficult to move, given that DNA is so widely-
trusted and reliable. On the other hand, confession, eyewitness, and fingerprint generally
had higher false alarm rates for the positive prime conditions than the negative prime
conditions. Perhaps if the study were more highly powered, the differences between
positive and negative priming conditions would become significant for confession,
eyewitness, and fingerprint, but not for DNA. Again, although this is purely speculative, the
results suggest that some forms of evidence yield different false recall results than others,
but, notably, to a very small degree.

Another potential explanation for why we did not observe differential activation
strength is that the semantic activation is not a reasoning judgment (“That person is
guilty”) but instead definitional (“Guilt exists in the criminal justice system”). Indeed,
“innocent” is one of our list words—the opposite of “guilty.” Participants may be activating
“guilty” in a way that is quite different from a juror making an assessment of a defendant.
After all, we are aware of no evidence that reading a DRM list affects a person’s judgments
AUTOMATICALLY GUILTY

or decision-making outside of that DRM task—for example, reading the “sleep” list does not make you sleepy, or decide to go to sleep. Study 3b provides weak evidence that supports the proposition that the semantic activation of “guilty” is different from determining whether someone is guilty. If participants are behaving in the decision-making manner of jurors, those in the negative prime conditions who read a vignette of a wrongfully convicted person should, activate “innocent” and suppress its opposite, “guilty.” Activating “innocent” should lead to lower false recall for “guilty” on the DRM rate. However, we found no main effect for Prime Type (nor significant differences between false recall rates for the negative and positive for each evidence type), suggesting that both negative and positive prime participants were processing the “guilty” list in the same way. Thus, perhaps participants that have cued “guilty” due to spreading activation are thinking of “guilty” in a qualitatively different way from jurors who are assessing guilt.

**Future Research**

Despite our lack of significant results, we believe there may yet be hope for using the DRM paradigm in this way. Each of our three studies yielded interesting trends, if not support for our hypotheses. Study 1 suggests that the forward associative strength of the list words may not always predict activation strength of the critical lure; and yet, our 7-word list still functioned very similarly to the 14-word list structure in terms of false alarms and confidence. In Study 2, the Story format yielded surprisingly high rates of Free Recall and Recognition test false recall for “guilty.” In Study 3, we found no main effects or interactions, but think that a more highly-powered study may yield significant differences between the positively- and negatively-primed evidence lists. As such, this project has raised a number of new questions that explore how the DRM paradigm works. For example,
we wonder if our “guilty” list functions differently from other DRM lists, given the
surprising effects of weaker-associated words (Study 1) and the high rates of false
memories and false alarms from the “guilty” Story (Study 2). Are participants processing
these criminal justice-related words in a way different from other, more concrete DRM
lists? Although we found our list functioned similarly to the “chair” list and “sleep” story
(although, in many cases, not identically; See Appendix D and G), future research should
investigate other aspects that differ between the “guilty” and other DRM lists and stories
that may account for our unexpected results.

The lack of priming effects was also surprising, given its success in past research
(Lenton et al., 2001; Takarangi et al., 2007). As noted previously, these studies primed the
entire content of the list, whereas we only primed the strength of association between one
word and the critical lure. We are curious whether single-word priming would affect
semantic activation for other, traditional DRM lists. Further, we suspect that a prime that
affects the entirety of the “guilty” list—perhaps using ambiguous guilt-related words
(Takarangi et al., 2007) would be more effective in altering how participants activate the
“guilty” network.

Given the applied nature of this set of studies, a logical next step for this paradigm is
assessing how it could be used in the real world. Although it is clear that further research is
necessary before it would be plausible to take this paradigm out of the lab, such future
research should focus on the specific applied situations for which it is suited—for example,
how this paradigm can be used as a tool with mock trial designs. We are interested in how
situational factors influence responses to the “guilty” DRM list and story. Do jurors respond
differently to the list after a prosecution’s case in chief than after a defense’s? Do individual
differences such as being a police officer have an effect? In some cases (Study 1b and Study 2b), we found contact with the criminal justice system changed our pattern of results. Given these comparisons were only a fraction of our required sample size, and significant differences are eliminated when combined with the overall sample, we are hesitant to interpret them as meaningful. However, these differences suggest contact with the criminal justice system may indeed be such an important individual difference. We need a better understanding of how the “guilty” DRM list and story works, but we think it has potential for real-world application.

Despite overall lack of support for our hypotheses that we could measure differential activation of “guilty” between different forms of evidence, our results do raise a number of theoretical and applied research questions for future research. We think that our novel use of the extensively-used and well-understood DRM paradigm lays the groundwork for a simple, effective tool for understanding a potential cause of wrongful conviction.
## Appendix A

### Forward Associative Task Results

<table>
<thead>
<tr>
<th>Word</th>
<th>List Position</th>
<th>Associative Strength</th>
<th>Word</th>
<th>List Position</th>
<th>Associative Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jury</td>
<td>1</td>
<td>47%</td>
<td>Jail</td>
<td>8</td>
<td>24%</td>
</tr>
<tr>
<td>Crime</td>
<td>2</td>
<td>44%</td>
<td>Murder*</td>
<td>9</td>
<td>17%</td>
</tr>
<tr>
<td>Judge</td>
<td>3</td>
<td>44%</td>
<td>Convict</td>
<td>10</td>
<td>20%</td>
</tr>
<tr>
<td>Court</td>
<td>4</td>
<td>39%</td>
<td>Shame</td>
<td>11</td>
<td>17%</td>
</tr>
<tr>
<td>Criminal</td>
<td>5</td>
<td>27%</td>
<td>Lawyer</td>
<td>12</td>
<td>15%</td>
</tr>
<tr>
<td>Innocent</td>
<td>6</td>
<td>24%</td>
<td>Verdict</td>
<td>13</td>
<td>10%</td>
</tr>
<tr>
<td>Trial</td>
<td>7</td>
<td>24%</td>
<td>Sentence</td>
<td>14</td>
<td>10%</td>
</tr>
</tbody>
</table>

*"Murder" was listed as the 9th word due to the fact that other specific crimes (robbery, stealing, etc.) were occasionally listed by participants.*
### Appendix B

**Exclusion Criteria**

<table>
<thead>
<tr>
<th>Question</th>
<th>Excluded if responded...</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Did you complete the experiment in a single session, without stopping?</td>
<td>No</td>
<td>Memory confound (unknown decay time between list and test)</td>
</tr>
<tr>
<td>2. Did you pause or leave the experiment to engage in other tasks?</td>
<td>Yes</td>
<td>Memory confound (unknown decay time between list and test)</td>
</tr>
<tr>
<td>3. Did you use your web browser’s back or refresh buttons during the experiment?</td>
<td>Yes</td>
<td>May have disrupted Qualtrics progression and assignment</td>
</tr>
<tr>
<td>4. Did you complete the experiment without anyone helping you?</td>
<td>No</td>
<td>Memory confound (memory aid, distraction from encoding, etc.)</td>
</tr>
<tr>
<td>5. Did you speak with anyone at any time during the experiment?</td>
<td>Yes</td>
<td>Memory confound (memory aid, distraction from encoding, etc.)</td>
</tr>
<tr>
<td>5. Did you use a search engine at any point to look anything up?</td>
<td>Yes</td>
<td>Memory confound (memory aid, distraction from encoding, etc.); in Study 3, potential interference with primes</td>
</tr>
<tr>
<td>6. Did you take notes?</td>
<td>Yes</td>
<td>Memory confound (memory aid)</td>
</tr>
<tr>
<td>7. Did you (intentionally or unintentionally) rewind or restart the video showing the list?*</td>
<td>Yes</td>
<td>Memory confound (multiple exposure to words)</td>
</tr>
<tr>
<td>8. Did you experience any of the following technical difficulties? Video Problems; Survey restarted; Other. Please Specify</td>
<td>Yes for video problems and survey restart; “Other” responses were assessed by the researcher</td>
<td>May interfere with Qualtrics group assignment; Video Problems may affect ability to remember words.</td>
</tr>
</tbody>
</table>

*Question 7 was not used as an exclusion rule for Study 2, as the Story was not presented in a video.*
Appendix C

Study 1a Word Lists

Guilty List
*Strong Association:* jury, crime, judge, court, criminal, innocent, trial
*Medium Association:* court, criminal, innocent, trial, jail, murder, convict
*Weak Association:* jail, murder, convict, shame, lawyer, verdict, sentence

*Medium-Strong:* crime, judge, court, criminal, innocent, trial, jail

Chair List
*Strong Association:* desk, cushion, couch, bench, sit, swivel, sofa
*Moderate Association:* bench, sit, swivel, sofa, recliner, rocking, sitting
*Weak Association:* recliner, rocking, sitting, legs, table, seat, wood

Study 1b Word Lists

Guilty List
*Control:* jury, crime, judge, court, criminal, innocent, trial
*Confession:* jury, crime, judge, confession, court, criminal, innocent
*DNA:* jury, crime, judge, DNA, court, criminal, innocent
*Eyewitness:* jury, crime, judge, eyewitness, court, criminal, innocent
*Fingerprint:* jury, crime, judge, fingerprint, court, criminal, innocent

*Guilty Lists containing evidence constructed by dropping “trial” and have the evidence word inserted as the fourth word.

Chair List
*Control:* desk, cushion, couch, bench, sit, swivel, sofa
### Appendix D

#### Study 1b “Guilty” and “Chair” List Descriptives and Comparisons

<table>
<thead>
<tr>
<th>Item Type</th>
<th>List</th>
<th>Free Recall (SD)</th>
<th>Recognition (SD)</th>
<th>Confidence (SD)</th>
<th>Remember (*)</th>
<th>Know (*)</th>
<th>Guess (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Lure</td>
<td>Guilty</td>
<td>.08 (.27)</td>
<td>.50 (.50)</td>
<td>3.73 (1.19)</td>
<td>174 (.344)</td>
<td>186 (369)</td>
<td>145 (287)</td>
</tr>
<tr>
<td>Chair</td>
<td>(Proportions)</td>
<td>.43 (.50)</td>
<td>.79 (.41)</td>
<td>3.98 (1.13)</td>
<td>399 (.498)</td>
<td>285 (355)</td>
<td>118 (147)</td>
</tr>
<tr>
<td>Old</td>
<td>Guilty</td>
<td>.91 (.18)</td>
<td>3.54 (0.84)</td>
<td>2.25 (.89)</td>
<td>.39 (.74)</td>
<td>.10 (.33)</td>
<td></td>
</tr>
<tr>
<td>Chair</td>
<td>(Counts)</td>
<td>.93 (.17)</td>
<td>4.62 (0.55)</td>
<td>2.26 (.90)</td>
<td>.42 (.71)</td>
<td>.11 (.35)</td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>Guilty</td>
<td>.03 (.12)</td>
<td>4.58 (0.69)</td>
<td>.01 (.14)</td>
<td>.02 (.18)</td>
<td>.04 (.27)</td>
<td></td>
</tr>
<tr>
<td>Chair</td>
<td>(Counts)</td>
<td>.04 (.13)</td>
<td>4.36 (0.90)</td>
<td>.02 (.16)</td>
<td>.03 (.20)</td>
<td>.06 (.30)</td>
<td></td>
</tr>
<tr>
<td>Weak Lure</td>
<td>Guilty</td>
<td>.31 (.46)</td>
<td>3.68 (1.26)</td>
<td>63 (201)</td>
<td>104 (.332)</td>
<td>146 (.466)</td>
<td></td>
</tr>
<tr>
<td>Chair</td>
<td>(Proportions)</td>
<td>.41 (.49)</td>
<td>3.53 (1.22)</td>
<td>84 (201)</td>
<td>171 (.410)</td>
<td>162 (.388)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes**

1. Bolded values are those that are significantly different (p<.05) between the two lists. Actual p-value and effect size (if applicable) reported below each set of values.
2. For Critical Lure and Weak Lure Free Recall and Recognition values, we report the proportion of participants that indicated they had seen the words on the list—either by including it in their Free Recall response, or by selecting “Old” in the Recognition Test.
3. For Old and New Word Recognition, we report the mean number of those words categorized as “Old” on the recognition test. Thus, higher values for Old words indicate higher accuracy in categorization; a score of 1.0 would signify categorizing all Old words as “Old.” Conversely, lower values for New words indicate higher accuracy in categorization; a score of 0 would signify categorizing all New words as “New.”
4. R/K/G for Critical Lure and Weak Lure, frequencies are reported (e.g., the number of participants who responded R/K/G) because participants could only respond with one value.
5. The value in parentheses for Critical Lure and Weak Lure report the percentage of each response type within all “Old” responses. These values were not used for any inferential statistics.
6. For R/K/G Old and New words, the recognition test contained three of each type (e.g., three old words, and three new words). Thus, we report the mean number of R responses, K responses, and G responses for each participant. Note that these values do not add up to 3 because only participants who labeled a word as "Old" made a R/K/G judgment.
7. The value in parentheses for Old and New Words represents the standard deviation for each value.
**Study 1 “Guilty” Placement Chi Square**

\[ X^2(8, N=71) = 13.00, p = .369. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>First Third</th>
<th>Middle Third</th>
<th>Last Third</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0\textsuperscript{a}</td>
<td>0\textsuperscript{a}</td>
<td>17\textsuperscript{a}</td>
<td>17</td>
</tr>
<tr>
<td>Confession</td>
<td>0\textsuperscript{a}</td>
<td>3\textsuperscript{a}</td>
<td>12\textsuperscript{a}</td>
<td>15</td>
</tr>
<tr>
<td>DNA</td>
<td>0\textsuperscript{a}</td>
<td>1\textsuperscript{a}</td>
<td>8\textsuperscript{a}</td>
<td>9</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>1\textsuperscript{a}</td>
<td>4\textsuperscript{a}</td>
<td>10\textsuperscript{a}</td>
<td>15</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>0\textsuperscript{a}</td>
<td>2\textsuperscript{a}</td>
<td>13\textsuperscript{a}</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>10</td>
<td>60</td>
<td>71</td>
</tr>
</tbody>
</table>

*Note: Matching superscript letters indicate categories whose column proportions do not differ significantly at \( p = .05 \) level.*
### Appendix E

**Study 1b t-Tests and Alternate Analyses**

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Free Recall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>.09 [.05, .13]</td>
<td></td>
<td>Confession</td>
<td>.08 [.04, .12]</td>
<td>t(404)=0.44</td>
<td>p=.662</td>
<td>d=.03 [.01, .07]</td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>.05 [.02, .08]</td>
<td></td>
<td></td>
<td>t(408)=1.71</td>
<td>p=.088</td>
<td>d=.16 [.13, .18]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td>.07 [.04, .11]</td>
<td></td>
<td></td>
<td>t(414)=0.65</td>
<td>p=.518</td>
<td>d=.07 [.05, .10]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>.08 [.04, .12]</td>
<td></td>
<td></td>
<td>t(396)=0.31</td>
<td>p=.755</td>
<td>d=.04 [.01, .06]</td>
</tr>
<tr>
<td>Confession</td>
<td>.08 [.04, .12]</td>
<td></td>
<td>DNA</td>
<td></td>
<td>t(402)=1.27</td>
<td>p=.205</td>
<td>d=.12 [.10, .15]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td></td>
<td></td>
<td></td>
<td>t(412)=0.20</td>
<td>p=.843</td>
<td>d=.04 [.01, .07]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(390)=0.12</td>
<td>p=.904</td>
<td>d=.00 [-.03, .03]</td>
</tr>
<tr>
<td>DNA</td>
<td>.05 [.02, .08]</td>
<td></td>
<td>EW</td>
<td></td>
<td>t(416)=1.09</td>
<td>p=.277</td>
<td>d=.08 [.06, .12]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(394)=1.39</td>
<td>p=.169</td>
<td>d=.12 [.10, .15]</td>
</tr>
<tr>
<td>EW</td>
<td>.07 [.04, .11]</td>
<td></td>
<td>FP</td>
<td></td>
<td>t(404)=0.32</td>
<td>p=.750</td>
<td>d=.04 [.01, .06]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Recognition (Old/New)</strong></th>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>.48 [.41, .55]</td>
<td></td>
<td>Confession</td>
<td>.53 [.46, .60]</td>
<td>t(404)=0.99</td>
<td>p=.321</td>
<td>d=-.10 [.05, .15]</td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>.46 [.39, .52]</td>
<td></td>
<td></td>
<td>t(408)=0.50</td>
<td>p=.617</td>
<td>d=.04 [-.01, .09]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td>.51 [.44, .58]</td>
<td></td>
<td></td>
<td>t(414)=.59</td>
<td>p=.557</td>
<td>d=.06 [.01, .11]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>.51 [.44, .58]</td>
<td></td>
<td></td>
<td>t(396)=.59</td>
<td>p=.553</td>
<td>d=.06 [.01, .11]</td>
</tr>
<tr>
<td>Confession</td>
<td>.53 [.46, .60]</td>
<td></td>
<td>DNA</td>
<td></td>
<td>t(402)=1.49</td>
<td>p=.137</td>
<td>d=.14 [.09, .19]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td></td>
<td></td>
<td></td>
<td>t(412)=.42</td>
<td>p=.675</td>
<td>d=.04 [-.01, .09]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(390)=.39</td>
<td>p=.403</td>
<td>d=.04 [-.01, .09]</td>
</tr>
<tr>
<td>DNA</td>
<td>.46 [.39, .52]</td>
<td></td>
<td>EW</td>
<td></td>
<td>t(416)=1.09</td>
<td>p=.275</td>
<td>d=.10 [.05, .15]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(394)=1.08</td>
<td>p=.279</td>
<td>d=.10 [.05, .15]</td>
</tr>
<tr>
<td>EW</td>
<td>.51 [.44, .58]</td>
<td></td>
<td>FP</td>
<td></td>
<td>t(404)=0.02</td>
<td>p=.983</td>
<td>d=.00 [-.05, .05]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Confidence</strong></th>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>3.83 [3.67, 3.99]</td>
<td></td>
<td>Confession</td>
<td>3.60 [3.43, 3.77]</td>
<td>t(404)=1.95</td>
<td>p=.052</td>
<td>d=.19 [.08, .31]</td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>3.80 [3.65, 3.96]</td>
<td></td>
<td></td>
<td>t(408)=.23</td>
<td>p=.817</td>
<td>d=.03 [-.08, .14]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td>3.70 [3.53, 3.86]</td>
<td></td>
<td></td>
<td>t(414)=1.15</td>
<td>p=.250</td>
<td>d=.11 [.00, .22]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>3.70 [3.53, 3.88]</td>
<td></td>
<td></td>
<td>t(396)=1.05</td>
<td>p=.293</td>
<td>d=.11 [-.01, .23]</td>
</tr>
<tr>
<td>Confession</td>
<td>3.60 [3.43, 3.77]</td>
<td></td>
<td>DNA</td>
<td></td>
<td>t(402)=1.77</td>
<td>p=.078</td>
<td>d=.17 [.06, .29]</td>
</tr>
<tr>
<td></td>
<td>EW</td>
<td></td>
<td></td>
<td></td>
<td>t(412)=.81</td>
<td>p=.418</td>
<td>d=.08 [-.03, .20]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(390)=.84</td>
<td>p=.403</td>
<td>d=.08 [-.04, .20]</td>
</tr>
<tr>
<td>DNA</td>
<td>3.80 [3.65, 3.96]</td>
<td></td>
<td>EW</td>
<td></td>
<td>t(416)=.95</td>
<td>p=.344</td>
<td>d=.09 [-.03, .20]</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td>t(394)=.86</td>
<td>p=.393</td>
<td>d=.08 [-.03, .20]</td>
</tr>
<tr>
<td>EW</td>
<td>3.70 [3.53, 3.86]</td>
<td></td>
<td>FP</td>
<td></td>
<td>t(404)=.06</td>
<td>p=.955</td>
<td>d=.00 [-.12, .12]</td>
</tr>
</tbody>
</table>
Study 1 Free Recall Chi-Square Alternate Analysis

\[ X^2(4, \, N=1016)=3.09, \, p=.543. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>Absent in FR</th>
<th>Present in FR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>184\textsuperscript{a}</td>
<td>16\textsuperscript{a}</td>
<td>200</td>
</tr>
<tr>
<td>Confession</td>
<td>187\textsuperscript{a}</td>
<td>19\textsuperscript{a}</td>
<td>206</td>
</tr>
<tr>
<td>DNA</td>
<td>194\textsuperscript{a}</td>
<td>10\textsuperscript{a}</td>
<td>204</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>198\textsuperscript{a}</td>
<td>16\textsuperscript{a}</td>
<td>214</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>176\textsuperscript{a}</td>
<td>16\textsuperscript{a}</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>939</td>
<td>77</td>
<td>1016</td>
</tr>
</tbody>
</table>

Note: Matching superscript letters indicate categories whose column proportions do not differ significantly at \( p=.05 \) level.

Study 1 Recognition Test Chi-Square Alternate Analysis

\[ X^2(4, \, N=1016)=2.74, \, p=.602. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>“New”</th>
<th>“Old”</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>107\textsuperscript{a}</td>
<td>99\textsuperscript{a}</td>
<td>206</td>
</tr>
<tr>
<td>Confession</td>
<td>94\textsuperscript{a}</td>
<td>106\textsuperscript{a}</td>
<td>200</td>
</tr>
<tr>
<td>DNA</td>
<td>111\textsuperscript{a}</td>
<td>93\textsuperscript{a}</td>
<td>204</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>105\textsuperscript{a}</td>
<td>109\textsuperscript{a}</td>
<td>214</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>94\textsuperscript{a}</td>
<td>98\textsuperscript{a}</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>511</td>
<td>505</td>
<td>1016</td>
</tr>
</tbody>
</table>

Note: Matching superscript letters indicate categories whose column proportions do not differ significantly at \( p=.05 \) level.
Study 2 DRM Story Format

**DRM Story Paragraph (14 words):**

A jury is a group of normal citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order and gives instructions. In some cases, family members and interested community members can be present in the court as well. If a person loses, they will be considered a criminal. If they win, they are usually innocent and show this by the end of the trial. If they lose, they are sent to jail, especially for murder. They are also labeled as a convict, which sometimes makes them feel shame. Of course, a person is not alone; they usually have a lawyer on their side to argue before the verdict, as well as help them get a fair sentence. Thus, many people are involved in the system.

**DRM Story Paragraph (7 words) Control:**

A jury is a group of normal citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order and gives instructions. In some cases, family members and interested community members can be present in the court as well. If a person loses, they will be considered a criminal. If they win, they are usually innocent.

**DRM Story Paragraph (7 words) evidence variations:**

A jury is a group of normal citizens that reviews information in a case to determine whether a crime has been committed. The judge provides order and gives instructions. They also make sure that evidence, such as _______ is presented. In some cases, family members and interested community members can be present in the court as well. If a person loses, they will be considered a criminal. If they win, they are usually innocent.

**Sleep Story (14 words):**

Sally lay in bed. She needed to rest but she was still awake even though she was so tired. Finally she nodded off and began to dream. She did not want to wake up. She began to snore and wrapped her blanket tightly around her. Sally quickly fell from a doze to a deep slumber and began to snore more heavily. Sally lay there in peace until she woke up and let out a big yawn. She was still drowsy.
Appendix G

Study 2b “Guilty” and “Sleep” Story Descriptives and Comparisons

<table>
<thead>
<tr>
<th>Item</th>
<th>Story</th>
<th>Free Recall (SD)</th>
<th>Recognition (SD)</th>
<th>Confidence (SD)</th>
<th>Remember (*)</th>
<th>Know (*)</th>
<th>Guess (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Lure</td>
<td>Guilty</td>
<td>.56 (.50)</td>
<td>.87 (.34)</td>
<td>4.40 (.96)</td>
<td>602 (.690)</td>
<td>219</td>
<td>52 (.060)</td>
</tr>
<tr>
<td>Sleep</td>
<td></td>
<td>.86 (.35)</td>
<td>.93 (.25)</td>
<td>4.48 (.85)</td>
<td>678 (.719)</td>
<td>230</td>
<td>35 (.037)</td>
</tr>
<tr>
<td>(Proportions)</td>
<td></td>
<td>p&lt;.000; d=.540</td>
<td>p&lt;.000; d=.189</td>
<td>p=.013; d=.078</td>
<td>McNemar-Bowker p&lt;.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td>Guilty</td>
<td>3.66 (1.68)</td>
<td>.95 (1.17)</td>
<td>4.68 (.56)</td>
<td>2.32 (.97)</td>
<td>.43 (.79)</td>
<td>.10 (.35)</td>
</tr>
<tr>
<td>Sleep</td>
<td></td>
<td>-.68 (.20)</td>
<td>3.94 (.68)</td>
<td>1.81 (1.2)</td>
<td>1.06</td>
<td>.55 (.81)</td>
<td></td>
</tr>
<tr>
<td>(Counts)</td>
<td></td>
<td>p&lt;.000; d=1.172</td>
<td>p&lt;.000; d=1.025</td>
<td>p=.000; d=.368</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>Guilty</td>
<td>-</td>
<td>.05 (.14)</td>
<td>4.42 (.62)</td>
<td>.03 (.17)</td>
<td>.04 (.21)</td>
<td>.08 (.32)</td>
</tr>
<tr>
<td>Sleep</td>
<td></td>
<td>-</td>
<td>.05 (.14)</td>
<td>4.35 (.71)</td>
<td>.07 (.31)</td>
<td>.11 (.39)</td>
<td>.16 (.519)</td>
</tr>
<tr>
<td>(Counts)</td>
<td></td>
<td>p=.523</td>
<td>p&lt;.001; d=.105</td>
<td>p&lt;.000; d=.117</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak Lure</td>
<td>Guilty</td>
<td>-</td>
<td>.42 (.49)</td>
<td>3.62 (1.19)</td>
<td>.75 (1.80)</td>
<td>186</td>
<td>156 (.374)</td>
</tr>
<tr>
<td>Sleep</td>
<td></td>
<td>-.85 (.36)</td>
<td>4.32 (1.01)</td>
<td>613 (.719)</td>
<td>178</td>
<td>62 (.073)</td>
<td></td>
</tr>
<tr>
<td>(Proportions)</td>
<td></td>
<td>p&lt;.000; d=.501</td>
<td>p&lt;.000; d=.540</td>
<td>McNemar-Bowker p&lt;.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes
1. Bolded values are those that are significantly different (p<.05) between the two lists. Actual p-value and effect size (if applicable) reported below each set of values.
2. For Critical Lure and Weak Lure Free Recall and Recognition values, we report the proportion of participants that indicated they had seen the words on the list—either by including it in their Free Recall response, or by selecting “Old” in the Recognition Test.
3. For Old and New Word Recognition, we report the mean number of those words categorized as “Old” on the recognition test. Thus, higher values for Old words indicate higher accuracy in categorization; a score of 1.0 would signify categorizing all Old words as “Old.” Conversely, lower values for New words indicate higher accuracy in categorization; a score of 0 would signify categorizing all New words as “New.”
4. R/K/G for Critical Lure and Weak Lure, frequencies are reported (e.g., the number of participants who responded R/K/G) because participants could only respond with one value.
5. The value in parentheses for Critical Lure and Weak Lure report the percentage of each response type within all “Old” responses. These values were not used for any inferential statistics.
6. For R/K/G Old and New words, the recognition test contained three of each type (e.g., three old words, and three new words for “guilty”; six and six for “sleep”). Thus, we report the mean number of R responses, K responses, and G responses for each participant. Note that these values do not add up to 3/6 because only participants who labeled a word as “Old” made a R/K/G judgment.
AUTOMATICALLY GUILTY

*7. The value in parentheses for Old and New Words represents the standard deviation for each value.

Study 2 “Guilty” Placement Chi Square

\[ X^2(8, N=566)=16.47, p=.036. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>First Third</th>
<th>Middle Third</th>
<th>Last Third</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>23\textsuperscript{a}</td>
<td>29\textsuperscript{ab}</td>
<td>53\textsuperscript{b}</td>
<td>114</td>
</tr>
<tr>
<td>Confession</td>
<td>17\textsuperscript{a}</td>
<td>29\textsuperscript{a}</td>
<td>68\textsuperscript{a}</td>
<td>105</td>
</tr>
<tr>
<td>DNA</td>
<td>15\textsuperscript{a}</td>
<td>30\textsuperscript{a}</td>
<td>86\textsuperscript{a}</td>
<td>131</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>13\textsuperscript{a}</td>
<td>22\textsuperscript{a}</td>
<td>65\textsuperscript{a}</td>
<td>100</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>7\textsuperscript{a}</td>
<td>26\textsuperscript{ab}</td>
<td>83\textsuperscript{b}</td>
<td>116</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>136</td>
<td>355</td>
<td>566</td>
</tr>
</tbody>
</table>

Note: Matching superscript letters indicate categories whose column proportions do not differ significantly at \( p=.05 \) level.
### Appendix H

**Study 2b t-Tests and Alternate Analyses**

#### Free Recall

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>.51 [.45, .58]</td>
<td>Confession</td>
<td>.55 [.49, .62]</td>
<td>t(404)=0.80</td>
<td>p=.423</td>
<td>d=.08 [.03, .13]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DNA</td>
<td>.66 [.59, .72]</td>
<td>t(403)=2.92</td>
<td>p=.004</td>
<td>d=.31 [.26, .35]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>.50 [.43, .57]</td>
<td>t(402)=0.30</td>
<td>p=.768</td>
<td>d=.02 [-.03, .07]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>.57 [.50, .59]</td>
<td>t(404)=1.10</td>
<td>p=.271</td>
<td>d=.12 [.07, .17]</td>
</tr>
<tr>
<td>Confession</td>
<td>.55 [.49, .62]</td>
<td>DNA</td>
<td>(401)=2.11</td>
<td>p=.036</td>
<td>d=.23 [.18, .27]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>t(400)=1.09</td>
<td>p=.275</td>
<td>d=.10 [.05, .15]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(402)=0.30</td>
<td>p=.764</td>
<td>d=.04 [-.01, .09]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>.66 [.59, .72]</td>
<td>EW</td>
<td>t(399)=3.21</td>
<td>p=.001</td>
<td>d=.33 [.28, .38]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(401)=1.80</td>
<td>p=.087</td>
<td>d=.17 [.14, .23]</td>
<td></td>
</tr>
<tr>
<td>EW</td>
<td>.50 [.43, .57]</td>
<td>FP</td>
<td>t(400)=1.39</td>
<td>p=.164</td>
<td>d=.14 [.09, .19]</td>
<td></td>
</tr>
</tbody>
</table>

#### Recognition (Old/New)

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>.84 [.79, .89]</td>
<td>Confession</td>
<td>.85 [.80, .90]</td>
<td>t(404)=0.23</td>
<td>p=.816</td>
<td>d=.03 [-.01, .06]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DNA</td>
<td>.89 [.84, .93]</td>
<td>t(403)=1.25</td>
<td>p=.214</td>
<td>d=.15 [.11, .18]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>.89 [.84, .93]</td>
<td>t(402)=1.23</td>
<td>p=.221</td>
<td>d=.15 [.11, .18]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>.86 [.81, .91]</td>
<td>t(404)=0.52</td>
<td>p=.605</td>
<td>d=.06 [.09, .02]</td>
</tr>
<tr>
<td>Confession</td>
<td>.85 [.80, .90]</td>
<td>DNA</td>
<td>t(401)=1.01</td>
<td>p=.313</td>
<td>d=.12 [.09, .15]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>t(400)=0.99</td>
<td>p=.322</td>
<td>d=.12 [.09, .15]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(402)=0.28</td>
<td>p=.777</td>
<td>d=.03 [-.01, .06]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>.89 [.84, .93]</td>
<td>EW</td>
<td>t(399)=0.02</td>
<td>p=.986</td>
<td>d=.00 [-.03, .03]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(401)=0.73</td>
<td>p=.467</td>
<td>d=.09 [.06, .12]</td>
<td></td>
</tr>
<tr>
<td>EW</td>
<td>.89 [.84, .93]</td>
<td>FP</td>
<td>t(400)=0.71</td>
<td>p=.478</td>
<td>d=.09 [.06, .12]</td>
<td></td>
</tr>
</tbody>
</table>

#### Confidence

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Group</th>
<th>Mean</th>
<th>t-test</th>
<th>p value</th>
<th>d value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>4.27 [4.13, 4.41]</td>
<td>Confession</td>
<td>4.39 [4.25, 4.52]</td>
<td>t(404)=1.18</td>
<td>p=.239</td>
<td>d=.12 [.02, .22]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DNA</td>
<td>4.45 [4.32, 4.58]</td>
<td>t(403)=1.80</td>
<td>p=.073</td>
<td>d=.18 [.08, .28]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>4.41 [4.28, 4.54]</td>
<td>t(402)=1.42</td>
<td>p=.156</td>
<td>d=.14 [.05, .24]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>4.47 [4.35, 4.59]</td>
<td>t(404)=2.09</td>
<td>p=.037</td>
<td>d=.21 [.14, .30]</td>
</tr>
<tr>
<td>Confession</td>
<td>4.39 [4.25, 4.52]</td>
<td>DNA</td>
<td>t(401)=0.65</td>
<td>p=.516</td>
<td>d=.15 [.11, .18]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EW</td>
<td>t(400)=0.25</td>
<td>p=.801</td>
<td>d=.06 [-.03, .16]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(402)=0.92</td>
<td>p=.358</td>
<td>d=.09 [.00, .18]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>4.45 [4.32, 4.58]</td>
<td>EW</td>
<td>t(399)=0.40</td>
<td>p=.691</td>
<td>d=.04 [-.05, .14]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
<td>t(401)=0.25</td>
<td>p=.806</td>
<td>d=.02 [-.07, .11]</td>
<td></td>
</tr>
<tr>
<td>EW</td>
<td>4.47 [4.35, 4.59]</td>
<td>FP</td>
<td>t(400)=0.66</td>
<td>p=.510</td>
<td>d=.07 [-.02, .16]</td>
<td></td>
</tr>
</tbody>
</table>
AUTOMATICALLY GUILTY

Study 2b Free Recall Chi-Square Alternate Analysis

\[ X^2(4, N=1009)=12.34, p=0.015. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>Absent in FR</th>
<th>Present in FR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>99&lt;sup&gt;a&lt;/sup&gt;</td>
<td>105&lt;sup&gt;a&lt;/sup&gt;</td>
<td>204</td>
</tr>
<tr>
<td>Confession</td>
<td>90&lt;sup&gt;a&lt;/sup&gt;</td>
<td>112&lt;sup&gt;a&lt;/sup&gt;</td>
<td>202</td>
</tr>
<tr>
<td>DNA</td>
<td>69&lt;sup&gt;a&lt;/sup&gt;</td>
<td>132&lt;sup&gt;b&lt;/sup&gt;</td>
<td>201</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>100&lt;sup&gt;a&lt;/sup&gt;</td>
<td>100&lt;sup&gt;a&lt;/sup&gt;</td>
<td>200</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>87&lt;sup&gt;a&lt;/sup&gt;</td>
<td>115&lt;sup&gt;a&lt;/sup&gt;</td>
<td>202</td>
</tr>
<tr>
<td>Total</td>
<td>445</td>
<td>564</td>
<td>1009</td>
</tr>
</tbody>
</table>

*Note: Matching subscript letters indicate categories whose column proportions do not differ significantly at p=.05 level.*

Study 2b Recognition Test Chi-Square Alternate Analysis

\[ X^2(4, N=1009)=2.59, p=0.629. \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>“New”</th>
<th>“Old”</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>32&lt;sup&gt;a&lt;/sup&gt;</td>
<td>172&lt;sup&gt;a&lt;/sup&gt;</td>
<td>204</td>
</tr>
<tr>
<td>Confession</td>
<td>30&lt;sup&gt;a&lt;/sup&gt;</td>
<td>172&lt;sup&gt;a&lt;/sup&gt;</td>
<td>202</td>
</tr>
<tr>
<td>DNA</td>
<td>23&lt;sup&gt;a&lt;/sup&gt;</td>
<td>178&lt;sup&gt;a&lt;/sup&gt;</td>
<td>201</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>23&lt;sup&gt;a&lt;/sup&gt;</td>
<td>177&lt;sup&gt;a&lt;/sup&gt;</td>
<td>200</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>28&lt;sup&gt;a&lt;/sup&gt;</td>
<td>174&lt;sup&gt;a&lt;/sup&gt;</td>
<td>202</td>
</tr>
<tr>
<td>Total</td>
<td>136</td>
<td>873</td>
<td>1009</td>
</tr>
</tbody>
</table>
Appendix I

Study 3b Evidence Prime Vignettes
We report here the final version of each vignette. These vignettes were piloted across three samples, with content added to each sample. The content added to the second version (Sample 2) is demarcated by an underline; the content added to the third version (Sample 3) is italicized.

Confession – Positive
Tom Watkins, of Milwaukee, WI, was recently convicted of first degree murder after a two-day trial. Police identified Watkins as a suspect in the murder of another Milwaukee man, Sam Perkins. After Perkins did not report for work for two days, police went to his house where they found his body. Police arrested Watkins after a preliminary investigation revealed a connection between Watkins and Perkins.

At trial, Watkins was convicted based primarily on his confession to the crime. The jury found Watkins’ confession to be very compelling evidence, because it contained many accurate details, and the police gathered it using correct, standard interrogation methods, making it very reliable and likely true. Further, the jury heard expert testimony that a person will rarely confess to a crime they did not commit. After the trial, the judge offered an opinion that a confession is reliable and is often correct, as it was in this case. The judge also wrote that confession evidence should be very convincing that the defendant committed the crime. Therefore, Watkins’ confession was key in making sure the right person was brought to justice.

Confession – Negative
Tom Watkins, of Milwaukee, WI, was recently exonerated after being wrongfully convicted of first degree murder based on improper evidence. Police identified Watkins as a suspect in the murder of another Milwaukee man, Sam Perkins. After Perkins did not report for work for two days, police went to his house where they found his body. Police arrested Watkins after a preliminary investigation revealed a connection between Watkins and Perkins.

At trial, Watkins was convicted based primarily on his confession to the crime. The jury found Watkins’ confession to be very compelling evidence, despite it containing many inaccurate details, and the police gathering it using coercive interrogation methods, making it unreliable and likely false. Further, the jury heard expert testimony that a person will often confess to a crime they did not commit. After the trial, the judge offered an opinion that a confession is not always reliable and can be incorrect, as it was in this case. The judge also wrote that confession evidence should not be completely convincing that the defendant committed the crime. Therefore, Watkins’ confession was key in his wrongful conviction.

DNA – Positive
After a three-day trial, Andrew Simpson was recently convicted of first degree murder. Police discovered the body of a man from Omaha, Nebraska, named Timothy Rico, in a dumpster behind a convenience store. The police identified Simpson as a person of interest after discovering he worked with Rico in the convenience store. After the police located Simpson, he was arrested and charged with the crime.

At trial, Simpson was convicted based primarily on his DNA being found on Rico’s body. The jury found the DNA to be very compelling evidence because the DNA sample was appropriately gathered, using the correct, standard methods, and analyzed accurately, making the match very reliable. Further, the jury heard expert testimony that DNA found on a body is rarely from someone other than the killer. After the trial, the judge offered an opinion that DNA evidence is reliable and is often
correct, as it was in this case. The judge also wrote that DNA evidence should be very convincing that the defendant committed the crime. Therefore, Simpson’s DNA was key in making sure the right person was brought to justice.

**DNA – Negative**

A judge recently overturned Andrew Simpson’s conviction, based on admittance of improper evidence. Previously, after a three-day trial, Andrew Simpson was wrongly convicted of first degree murder. Police discovered the body of a man from Omaha, Nebraska, named Timothy Rico, in a dumpster behind a convenience store. The police identified Simpson as a person of interest after discovering he worked with Rico in the convenience store. After the police located Simpson, he was arrested and charged with the crime.

At trial, Simpson was convicted based primarily on flawed evidence that his DNA was found on Rico’s body. The jury found the DNA to be very compelling evidence, despite it being incorrectly gathered, and analyzed using an untested method that was shown to produce inaccuracies, making the match very unreliable. Further, the jury heard expert testimony that DNA found on a body is commonly from someone other than the killer. *After the trial, the judge offered an opinion that DNA is not always reliable and can be incorrect, as it was in this case. The judge also wrote that DNA evidence should not be completely convincing that the defendant committed the crime. Therefore, Simpson’s DNA was key in his wrongful conviction.*

**Eyewitness – Positive**

A Kansas City man named Robert Keller was convicted of first degree murder this week. Neighbors discovered the body of Rick Alleto in his truck, parked outside his own home, apparently the victim of a gunshot wound to the head. After speaking with neighbors, the police arrested Keller on suspicion of murder. After a week-long investigation, the prosecutor formally charged Keller with the homicide.

At trial, Keller was convicted based primarily on testimony from an eyewitness who said he saw Keller leaving Alleto’s house. The jury found the eyewitness’s identification to be very compelling evidence, because the police used proper, unbiased lineup procedures, and the eyewitness expressed high confidence in his identification, making it very reliable. Further, the jury heard expert testimony that under the circumstances surrounding Alleto’s death, the eyewitness had a good view of the defendant and would be able to make an accurate identification. *After the trial, the judge offered an opinion that an eyewitness is reliable and is often correct, as it was in this case. The judge also wrote that eyewitness evidence should be very convincing that the defendant committed the crime. Therefore, the eyewitness testimony was key in making sure the right person was brought to justice.*

**Eyewitness - Negative**

A Kansas City man named Robert Keller was exonerated this week after being wrongfully convicted of first degree murder based on improper evidence. Neighbors discovered the body of Rick Alleto in his truck, parked outside his own home, apparently the victim of a gunshot wound to the head. After speaking with neighbors, the police arrested Keller on suspicion of murder. After a week-long investigation, the prosecutor formally charged Keller with the homicide.

At trial, Keller was convicted based primarily on testimony from an eyewitness who said he saw Keller leaving Alleto’s house. The jury found the eyewitness’s identification to be very compelling evidence, despite the police having used faulty, contaminated lineup procedures, and the eyewitness expressing low confidence in his identification, making it very unreliable. Further, the jury heard expert testimony that under the circumstances surrounding Alleto’s death, the eyewitness had a poor
view of the defendant and would be unable to make an accurate identification. After the trial, the judge offered an opinion that an eyewitness is not always reliable and can be incorrect, as it was in this case. The judge also wrote that eyewitness evidence should not be completely convincing that the defendant committed the crime. Therefore, the eyewitness testimony was key in Keller’s wrongful conviction.

**Fingerprint – Positive**

After a 10-day trial, Mark Hampton of Gary, Indiana, was convicted of first degree murder. A month after being reported missing by his son, the police found the body of Samuel Lithgow in a Gary, Indiana landfill. The police uncovered forensic evidence of Hampton’s involvement, and the state prosecutor charged Hampton with the crime.

At trial, Hampton was convicted based primarily on his fingerprint being found on Lithgow’s body. The jury found the fingerprint to be very compelling evidence, because it was collected using standard methods, under pristine conditions, and analyzed by an experienced forensic examiner, leading to a very reliable match. Further, the jury heard expert testimony that fingerprints found on a body are rarely from someone other than the killer. After the trial, the judge offered an opinion that fingerprinting is reliable and is often correct, as it was in this case. The judge also wrote that a fingerprint match evidence should be very convincing that the defendant committed the crime. Therefore, Hampton’s fingerprint was key in making sure the right person was brought to justice.

**Fingerprint – Negative**

After a second trial, a judge overturned the conviction of Mark Hampton of Gary, Indiana. Previously, Hampton was wrongfully convicted of first degree murder based on improper evidence. A month after being reported missing by his son, the police found the body of Samuel Lithgow in a Gary, Indiana landfill. The police uncovered forensic evidence of Hampton’s involvement, and the state prosecutor charged Hampton with the crime.

At trial, Hampton was convicted based primarily on his fingerprint being found on Lithgow’s body. The jury found the fingerprint to be very compelling evidence, despite it being collected using incorrect methods, under conditions that degraded the sample, and analyzed by a forensic examiner still in training, leading to an inaccurate match. Further, the jury heard expert testimony that fingerprints found on a body are often from someone other than the killer. After the trial, the judge offered an opinion that fingerprinting is not always reliable and can be incorrect, as it was in this case. The judge also wrote that a fingerprint should not be completely convincing that the defendant committed the crime. Therefore, Hampton’s fingerprint was key in his wrongful conviction.

**Filler Paragraphs**

1.
Wally Taylor, at the age of 55, began experiencing extreme headaches, blackouts, migraines, and trouble concentrating. He first noticed these symptoms when he had trouble performing his job. After approximately two months of these symptoms, he was rushed to the hospital after blacking out at his daughter’s track meet. There, he met with a neurologist.

During his hospital visit, Taylor’s neurologist ran a number of medical tests including MRI and CT scans. Taylor was kept in the hospital for one night for observation, but quickly released because preliminary tests showed no obvious cause for the blackouts. After another month of testing, the neurologist diagnosed Taylor with a small brain tumor. The neurologist concluded that because the tumor was so small, it was difficult to find on the brain scans. The neurologist also recommended that surgery would be extremely effective, as the tumor was so small and had not yet spread to any other
areas of the brain. With surgery and a short round of chemotherapy, Wally Taylor will likely make a remarkably fast and successful recovery.

2. Gregory Sanford is a successful chef at a popular Minneapolis, MN steak restaurant. He has won many awards for his cooking abilities and received international recognition for his innovative recipes. Sanford was hired to his current job in 2012, after the restaurant's owner read an article about Sanford and his creative fusion of Caribbean and Americana cuisine.

This year, a high-profile business CEO hated one of Sanford’s new dishes and complained about the meal both on the internet and directly to the restaurant owner. The athlete complained that the meal was overcooked and not properly seasoned, and that the quality of the meat and vegetables in the meal was extremely low. The manager concluded that Sanford was at fault for the terrible meal, and fired him. The manager reasoned that Sanford should have paid extra care to the athlete’s meal. Further, Sanford is ultimately responsible for the quality of the food the restaurant orders, so if the food is low quality it is because Sanford was either cutting costs or not monitoring the orders. Although the owner provided Sanford severance pay, the owner has stated he would not recommend Sanford for any other jobs.
### Appendix J

#### Study 3b “Guilty” List Descriptives

<table>
<thead>
<tr>
<th>Item Type</th>
<th>List</th>
<th>Free Recall (SD)</th>
<th>Recognition (SD)</th>
<th>Confidence (SD)</th>
<th>Remember (*)</th>
<th>Know (*)</th>
<th>Guess (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Lure</td>
<td>Guilty</td>
<td>.11 (.31)</td>
<td>.61 (.49)</td>
<td>3.57 (1.18)</td>
<td>143 (.360)</td>
<td>132</td>
<td>122</td>
</tr>
<tr>
<td>Old</td>
<td>Guilty</td>
<td>4.49 (1.40)</td>
<td>.85 (.19)</td>
<td>3.28 (.47)</td>
<td>2.01 (.83)</td>
<td>.39 (.67)</td>
<td>.17 (.41)</td>
</tr>
<tr>
<td>New</td>
<td>Guilty</td>
<td>.82 (.37)</td>
<td>.03</td>
<td>(.14)</td>
<td>.02 (.19)</td>
<td>.04 (.23)</td>
<td>.04 (.25)</td>
</tr>
<tr>
<td>Weak Lure</td>
<td>Guilty</td>
<td>-.37 (.48)</td>
<td>3.55 (1.21)</td>
<td>33 (.134)</td>
<td>86 (.351)</td>
<td>126</td>
<td>(.514)</td>
</tr>
</tbody>
</table>

**Notes**

1. For Critical Lure and Weak Lure Free Recall and Recognition values, we report the proportion of participants that indicated they had seen the words on the list—either by including it in their Free Recall response, or by selecting “Old” in the Recognition Test.
2. For Old and New Word Recognition, we report the mean number of those words categorized as “Old” on the recognition test. Thus, higher values for Old words indicate higher accuracy in categorization; a score of 1.0 would signify categorizing all Old words as “Old.” Conversely, lower values for New words indicate higher accuracy in categorization; a score of 0 would signify categorizing all New words as “New.”
3. R/K/G for Critical Lure and Weak Lure, frequencies are reported (e.g., the number of participants who responded R/K/G) because participants could only respond with one value.
4. The value in parentheses for Critical Lure and Weak Lure report the percentage of each response type within all “Old” responses. These values were not used for any inferential statistics.
5. For R/K/G Old and New words, the recognition test contained three of each type (e.g., three old words, and three new words). Thus, we report the mean number of R responses, K responses, and G responses for each participant. Note that these values do not add up to 3 because only participants who labeled a word as “Old” made a R/K/G judgment.
6. The value in parentheses for Old and New Words represents the standard deviation for each value.
**Study 2 “Guilty” Placement Chi Square**

\[ X^2(24, N=72)=31.19, \ p=.148. \] (Overall Chi Square)

<table>
<thead>
<tr>
<th>Condition</th>
<th>First Third</th>
<th>Middle Third</th>
<th>Last Third</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0(^a)</td>
<td>2(^a)</td>
<td>7(^a)</td>
<td>9</td>
</tr>
<tr>
<td>Confession-</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>5(^a)</td>
<td>6</td>
</tr>
<tr>
<td>Confession</td>
<td>0(^a)</td>
<td>3(^a)</td>
<td>1(^a)</td>
<td>4</td>
</tr>
<tr>
<td>Confession+</td>
<td>0(^a)</td>
<td>0(^a)</td>
<td>5(^a)</td>
<td>5</td>
</tr>
<tr>
<td>DNA-</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>1(^a)</td>
<td>2</td>
</tr>
<tr>
<td>DNA</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>4(^a)</td>
<td>5</td>
</tr>
<tr>
<td>DNA+</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>5(^a)</td>
<td>6</td>
</tr>
<tr>
<td>Eyewitness-</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>8(^a)</td>
<td>9</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>3(^a)</td>
<td>4</td>
</tr>
<tr>
<td>Eyewitness+</td>
<td>0(^a)</td>
<td>2(^a)</td>
<td>3(^a)</td>
<td>5</td>
</tr>
<tr>
<td>Fingerprint-</td>
<td>0(^a)</td>
<td>0(^a)</td>
<td>9(^a)</td>
<td>9</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1(^a)</td>
<td>1(^a)</td>
<td>2(^a)</td>
<td>4</td>
</tr>
<tr>
<td>Fingerprint+</td>
<td>0(^a)</td>
<td>1(^a)</td>
<td>3(^a)</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1</td>
<td>15</td>
<td>56</td>
<td>72</td>
</tr>
</tbody>
</table>

*Note: Matching superscript letters indicate categories whose column proportions do not differ significantly at \( p=.05 \) level.*

\[ X^2(4, N=72)=4.99, \ p=.288. \] (Priming Condition Chi Square)

\[ X^2(8, N=72)=4.29, \ p=.830. \] (Evidence List Chi Square)
**Appendix K**

**Study 3b t-Tests and Alternate Analyses**

**Free Recall**

<table>
<thead>
<tr>
<th>List</th>
<th>Prime</th>
<th>Mean [95% CI]</th>
<th>Positive-Negative comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td></td>
<td>.18 [.07, .28]</td>
<td></td>
</tr>
<tr>
<td>Confession</td>
<td>Neutral</td>
<td>.08 [.00, .16]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.10 [.01, .19]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.12 [.03, .21]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Neutral</td>
<td>.10 [.01, .18]</td>
<td>$t(98)=0.32$, $p=.752$ $d=.06$, [.00, .13]</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.10 [.01, .19]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.04 [-.02, .10]</td>
<td></td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Neutral</td>
<td>.06 [-.01, .13]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.10 [.01, .19]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.18 [.07, .29]</td>
<td></td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Neutral</td>
<td>.06 [-.01, .13]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.08 [.00, .15]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.18 [.09, .27]</td>
<td></td>
</tr>
</tbody>
</table>

**Recognition (Old/New)**

<table>
<thead>
<tr>
<th>List</th>
<th>Prime</th>
<th>Mean [95% CI]</th>
<th>Positive-Negative Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td></td>
<td>.57 [.43, .71]</td>
<td></td>
</tr>
<tr>
<td>Confession</td>
<td>Neutral</td>
<td>.60 [.46, .74]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.74 [.61, .87]</td>
<td>$t(98)=1.90$, $p=.060$ $d=.39$, [.29, .48]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.56 [.42, .70]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Neutral</td>
<td>.67 [.53, .80]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.62 [.48, .76]</td>
<td>$t(98)=0.61$, $p=.547$ $d=.12$, [.02, .22]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.56 [.42, .70]</td>
<td></td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Neutral</td>
<td>.56 [.42, .70]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.60 [.46, .74]</td>
<td>$t(98)=0.80$, $p=.425$ $d=.16$, [.07, .26]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.52 [.38, .66]</td>
<td></td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Neutral</td>
<td>.65 [.51, .78]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>.61 [.47, .75]</td>
<td>$t(99)=0.33$, $p=.742$ $d=.06$, [-.03, .16]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>.64 [.50, .78]</td>
<td></td>
</tr>
</tbody>
</table>


Confidence

<table>
<thead>
<tr>
<th>List</th>
<th>Prime</th>
<th>Mean [95% CI]</th>
<th>Positive-Negative comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Neutral</td>
<td>3.84 [3.56, 4.12]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3.44 [3.14, 3.74]</td>
<td>t(98)=0.36 p=.719 d=.07 [-.14, .29]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3.52 [3.19, 3.85]</td>
<td></td>
</tr>
<tr>
<td>Confession</td>
<td>Neutral</td>
<td>3.84 [3.56, 4.12]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3.44 [3.14, 3.74]</td>
<td>t(98)=0.36 p=.719 d=.07 [-.14, .29]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3.52 [3.19, 3.85]</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Neutral</td>
<td>3.49 [3.18, 3.81]</td>
<td>t(98)=1.55 p=.123 d=.31 [.09, .54]</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3.74 [3.40, 4.08]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3.38 [3.06, 3.70]</td>
<td></td>
</tr>
<tr>
<td>Eyewitness</td>
<td>Neutral</td>
<td>3.62 [3.27, 3.97]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3.32 [2.95, 3.69]</td>
<td>t(98)=1.56 p=.122 d=.32 [.07, .56]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3.72 [3.37, 4.07]</td>
<td></td>
</tr>
<tr>
<td>Fingerprint</td>
<td>Neutral</td>
<td>3.41 [3.07, 3.75]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>3.32 [2.96, 3.71]</td>
<td>t(99)=1.20 p=.234 d=.24 [.01, .48]</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>3.62 [3.48, 3.66]</td>
<td></td>
</tr>
</tbody>
</table>

Study 3b Free Recall Logistic Regression Alternate Analysis

$X^2(4, N=654)=7.16, p=.306.$

Nagelkerke $R^2=.022$

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>No Prime</td>
<td></td>
<td></td>
<td>3.503</td>
<td>2</td>
<td>.173</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Prime</td>
<td>.263</td>
<td>.361</td>
<td>3.50</td>
<td>1</td>
<td>.467</td>
<td>1.301</td>
<td>.641</td>
</tr>
<tr>
<td>Negative Prime</td>
<td>.622</td>
<td>.341</td>
<td>3.328</td>
<td>1</td>
<td>.068</td>
<td>1.863</td>
<td>.955</td>
</tr>
<tr>
<td>Control List</td>
<td></td>
<td></td>
<td>5.626</td>
<td>4</td>
<td>.229</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confession</td>
<td>-.978</td>
<td>.513</td>
<td>3.635</td>
<td>1</td>
<td>.057</td>
<td>.376</td>
<td>.138</td>
</tr>
<tr>
<td>DNA</td>
<td>-1.230</td>
<td>.529</td>
<td>5.411</td>
<td>1</td>
<td>.020</td>
<td>.292</td>
<td>.104</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>-.837</td>
<td>.505</td>
<td>2.747</td>
<td>1</td>
<td>.097</td>
<td>.433</td>
<td>.161</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>-.918</td>
<td>.508</td>
<td>3.262</td>
<td>1</td>
<td>.071</td>
<td>.399</td>
<td>.147</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.540</td>
<td>.367</td>
<td>17.588</td>
<td>1</td>
<td>.000</td>
<td>.214</td>
<td></td>
</tr>
</tbody>
</table>
**Study 3b Recognition Test Logistic Regression Alternate Analysis**

\[ X^2(4, N=654)=4.82, p=.567. \]

Nagelkerke \( R^2 = 0.010 \)

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Prime</td>
<td></td>
<td></td>
<td>2.261</td>
<td>2</td>
<td>.323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Prime</td>
<td>.099</td>
<td>.207</td>
<td>.230</td>
<td>1</td>
<td>.632</td>
<td>1.104</td>
<td>.736 - 1.656</td>
</tr>
<tr>
<td>Negative Prime</td>
<td>-.203</td>
<td>.204</td>
<td>.991</td>
<td>1</td>
<td>.320</td>
<td>.816</td>
<td>.548 - 1.217</td>
</tr>
<tr>
<td>Control List</td>
<td></td>
<td></td>
<td>2.657</td>
<td>4</td>
<td>.617</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confession</td>
<td>.307</td>
<td>.351</td>
<td>.767</td>
<td>1</td>
<td>.381</td>
<td>1.359</td>
<td>.684 - 2.702</td>
</tr>
<tr>
<td>DNA</td>
<td>.232</td>
<td>.349</td>
<td>.441</td>
<td>1</td>
<td>.506</td>
<td>1.261</td>
<td>.636 - 2.501</td>
</tr>
<tr>
<td>Eyewitness</td>
<td>.000</td>
<td>.348</td>
<td>.000</td>
<td>1</td>
<td>.999</td>
<td>1.000</td>
<td>.506 - 1.979</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>.298</td>
<td>.350</td>
<td>.727</td>
<td>1</td>
<td>.394</td>
<td>1.347</td>
<td>.679 - 2.675</td>
</tr>
<tr>
<td>Constant</td>
<td>.276</td>
<td>.283</td>
<td>.955</td>
<td>1</td>
<td>.329</td>
<td>1.318</td>
<td></td>
</tr>
</tbody>
</table>
Appendix L

DRM Instructions and Memory Measures

**DRM Primary Instructions:**

*Instructions (Beginning of survey):* For this task, we are going to have you read several [list of words]/[paragraphs] and ask you about them. Please do not write down anything from the [lists]/[paragraphs], take notes, etc. – instead, remember the [lists]/[paragraphs] as best you can.

*Instructions (List, Study 1a, 1b, and 3b):* Now we are going to have you read a list of words. We want you to remember the words as best you can. When you ready to begin, press “Next” and click the arrow on the video to begin the video.

*Instructions (Story, Study 2a and 2b):* Now, you will read a short paragraph. Again, please do not write down anything in the paragraph.

When you are ready to read the paragraph, proceed to the next page. You will have up to two minutes to read the paragraph, at which point you will be automatically advanced.

**Free Recall Test (Study 1a, 1b and Study 3b)**

*Instructions:* Now, we will test your memory for the list you just read. You will have 2 minutes to enter as many words as you’re reasonably confident you saw on the list. After 2 minutes, the survey will advance to the next part.

[Participants are given 25 blanks to enter words they remember]

**Free Recall Test (Study 2a and 2b)**

*Instructions:* Below, please recall the paragraph as accurately as you can. To the best of your ability, replicate the paragraph word-for-word, or as closely as you can.

[Participants are given an Essay Text Box to enter their response]

**Recognition Test**
You will now be asked some questions about the list you saw. We are testing your memory for the list.

Each question has two parts:
1) the first part asks you about a particular word from the list;
2) the second part asks you how confident you are with your answer.

Here is a sample question.

Experiment

Select "Old" if the word appeared on the previous list. If it did not appear on the list, and this is the first time you're seeing the word, select "New."

Old

New

(continued on next page)
AUTOMATICALLY GUILTY

If you select "Old" to indicate you did see the word on the list, you will then be asked the following:

Please indicate whether you Remember the word (you can recall specific details about seeing it and remembering it), if you Know the word (you have a sense of familiarity, but cannot recall any details), or if you just Guessed.

A "Remember" judgment indicates you can specifically remember seeing the word and have memories for the details, such as where in the list it appeared or what you thought about when you saw it.

A "Know" judgment indicates that while you cannot remember any of these types of details, you have a general sense of familiarity for the word. For example, you have a feeling the word appeared, even if you cannot specifically remember seeing it.

Finally, a "Guess" judgment indicates that your response was merely a guess.

- Remember
- Know
- Guess

How confident are you that your answer is correct?

Not at all confident  
Very confident

WHEN YOU HAVE READ AND UNDERSTOOD HOW TO ANSWER THESE QUESTIONS, CLICK NEXT TO BEGIN THE TEST.

(continued on next page)
Participants are then presented with the following words:

Judge\textsuperscript{2} Zebra\textsuperscript{3}
Balance\textsuperscript{3} Blossom\textsuperscript{3}
Court\textsuperscript{2} Develop\textsuperscript{3}
Criminal\textsuperscript{2} Castle\textsuperscript{3}
Guilty\textsuperscript{1} Jury\textsuperscript{2}
Trial\textsuperscript{2} Punishment\textsuperscript{4}
Elevator\textsuperscript{3}

Note: Order of words on the memory test was randomized, but the same across participants.

Key:
\textsuperscript{1} Critical Lure
\textsuperscript{2} Old Word (word presented on the list)
\textsuperscript{3} New Word (random word, not on the list)
\textsuperscript{4} Weakly Associated Lure (the 15\textsuperscript{th} word from our backwards association task; used as a theme-relevant control)
References


Brainerd, C. J., & Reyna, V. F. (1998). When things that were never experienced are easier to “remember” than things that were. Psychological Science, 9, 484-489.

AUTOMATICALLY GUILTY


AUTOMATICALLY GUILTY


AUTOMATICALLY GUILTY


AUTOMATICALLY GUILTY


