



City Research Online

City St George's, University of London

Citation: Elmore, T.L. (2020). The effect of cued memory recognition strategies on word and speaker identification. (Unpublished Doctoral thesis, City, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/25718/>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

The Effect of Cued Memory Recognition Strategies on Word and
Speaker Identification

Tiffany Lauren Elmore

February, 2020

A thesis submitted for the degree of Doctor of Philosophy (PhD) of City, University
of London, Department of Psychology

Table of Contents

Chapter 1	9
1.1 Introduction	9
1.2 Eyewitness Identification is Flawed.....	10
1.3 What is Earwitness Identification?.....	11
1.4 Earwitness Identification in History.....	13
1.5 Who is an Earwitness?	14
1.6 Confidence	15
1.7 Earwitness Memory.....	16
1.8 Legal Implications.....	17
1.9 Chapter Summaries	18
Chapter 2 – Literature Review	22
2.1 Factors that Affect an Earwitness’ Identification Performance	23
2.2 Factors that Impact a Listener’s Memory	37
2.3 Forensic Implications	45
2.4 Literature Summary.....	47
Chapter 3 – Experiment 1, Pilot Studies 1 and 2, Experiment 2	50
3.1 Introduction	50
3.2 Experiment 1	57
3.2.1 Method	58
3.2.2 Results.....	61
3.3 Pilot Study 1	64
3.3.1 Method	64
3.3.2 Results.....	67
3.4 Pilot Study 2	68
3.4.1 Method	69
3.4.2 Results.....	70
3.5 Experiment 2	72
3.5.1 Method	73
3.5.2 Results.....	75
3.5.3 Discussion	77
Chapter 4 – Experiment 3	83
4.1 - Introduction.....	83

4.1.1	Method.....	86
4.1.2	Results	88
4.1.3	Discussion.....	90
Chapter 5 – Experiment 4.....		93
5.1	Introduction.....	93
5.1.1	Method.....	94
5.1.2	Results	97
5.1.3	Discussion.....	99
Chapter 6 – Experiment 5.....		102
6.1	Introduction.....	102
6.1.1	Method.....	104
6.1.2	Results	106
6.1.3	Discussion.....	110
Chapter 7 – Experiment 6.....		112
7.1	Introduction.....	112
7.1.1	Method.....	115
7.1.2	Results	118
7.1.3	Discussion.....	121
Chapter 8 - General Discussion.....		125
8.1	Core Research Findings	125
8.2	Practical Limitations	136
8.3	Suggestions for Future Research	139
Appendices		144
References		186

List of Tables

Table 1-1: Overview of six experiments.....	20
Table 2-1: Overview of factors that impact speaker identification performance.....	49
Table 3-1: Mean c response bias and d' scores for Auditory Word Recognition.....	62
Table 3-2: Mean c response bias and d' scores for Speaker Identification.....	63
Table 3-3: Mean d' scores for Voice Identification in Repeated Exposure.....	67
Table 3-4: Mean d' scores for Voice Identification in Repeated Voice Exposure....	71
Table 3-5: Mean d' scores for Voice Identification in Male and Female Robber condition.....	76
Table 4-1: Mean d' scores for Statement Content Recognition.....	89
Table 5-1: Mean d' scores for Written and Auditory Statement Recognition.....	98
Table 6-1: Mean d' scores for Speaker Identification and mean scores for Familiarity and Confidence ratings.....	106

List of Figures

Figure 3-1: Mean d' scores for auditory word recognition for speaker gender.....	62
Figure 3-2: Mean d' scores for speaker identification based on speaker gender.....	64
Figure 3-3: Mean d' scores for speaker identification based on speaker gender	68
Figure 3-4: Mean d' scores for speaker identification based on speaker gender	72
Figure 3-5: Mean d' scores for memory voice identification for Robber gender as a function of speaker gender.....	77
Figure 4-1: Mean d' scores for content recognition by speaker gender as a function of duration.....	90
Figure 5-1: Mean d' scores for written and auditory statements recognition for speaker gender.....	99
Figure 6-1: ROC curve for Speaker Identification Accuracy.....	108
Figure 6-2: ROC curve for Speaker Identification Accuracy and Familiarity.....	109
Figure 6-3: ROC curve for Speaker Identification Accuracy and Confidence.....	109
Figure 7-1: Mean recall idea unit scores for participant gender as a function of speaker gender	119
Figure 7-2: Mean recall idea unit scores for participant gender as a function of crime scenario	121

Acknowledgements

I would like to extend my sincerest appreciation and gratitude to my wonderful supervisors, Joanne Rechdan and Jessica Jones Nielsen. With your passion and support, I was able to overcome many obstacles. I can't thank you enough for guiding me on this journey.

Many thanks to the dedicated staff, co-workers, and colleagues in the Psychology Department who shared their kindness and support to help me in my budding academic career.

I want to thank my family who call me everyday to support me and encourage me while I pursue my dreams. I could not be here without your love and compassion.

Finally, to my loving husband who met me just three months into this whole process, thank you for being you. You are my best friend and my rock. I am eternally blessed to have you in my life.

Declaration

I grant powers of discretion to the Director of Library Services to copy this thesis in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

Abstract

Previous research suggests that earwitness identification is flawed due to suggestive lineup techniques, poor witness memory, and challenges presented during and after the initial voice exposure. Earwitness evidence presented during court testimony is given substantial weight by jurors (Semmler, Brewer, & Douglass, 2012). The reliability of earwitness evidence is an understudied issue compared to eyewitness identification and warrants further exploration. To address the disparity in research, this thesis explored: (1) how well witnesses remember voices, (2) does speaker identification accuracy vary with the gender of the speaker, (3) does speaker identification accuracy vary when the witness is presented with a new voice or new phrase, (4) does speaker familiarity or confidence ratings predict speaker identification accuracy, and (5) how well witnesses recall details of a crime.

The overall aim of this thesis was to investigate the boundary condition for accurate recognition of voices and recall of verbal content. This was addressed in six experiments. The six experiments focused specifically on speaker identification accuracy. In Experiment 1, we evaluated how well witnesses remembered words and voices. In Experiments 3 and 4, we assessed whether exposure duration (Exp. 3) and source confusion (Exp. 4) would impact the encoding of written and auditory statements. In Experiment 5, we did not find that participants' familiarity or confidence ratings predicted speaker identification accuracy. In Experiment 2, we analyzed how well participants recognized voices associated with a criminal incident and found that overall, speaker identification accuracy was poor. The information that we gathered from our research has shown that memory for speaker identification is poor even when tested within controlled laboratory conditions.

Finally, to further contribute to reducing earwitness identification inaccuracies, we created a mobile application for recording and reporting important information. In Experiment 6, we reasoned that capturing crime-related information in real-time or immediately after an event would help to reduce memory errors that tend to increase with the passage of time (Yarmey and Matthys, 1992; Öhman, Eriksson, & Granhag, 2013). Such a tool will hopefully increase public safety and reduce eyewitness errors by serving as a technological corroborator.

Chapter 1

1.1 Introduction

On October 2, 1971, a young female nurse was approached by a gunman in a hospital parking lot and forced into her car. The gunman made her drive to a secluded area where he sexually assaulted her and ran away. The victim later gave three conflicting estimates of the assailant's height while also mentioning that he had a gap in his front teeth and spoke with a "smooth, soft voice." A few months later, a suspect provided an alibi that he was with a man named Wilbert Jones when the attempted rape occurred. The police apprehended Jones and, later, included him in a lineup viewed by the victim.

Each person in the lineup spoke a phrase similar to a statement made by the victim's assailant during the night of the attack. The victim picked Jones out of the lineup and identified him as her assailant. She expressed her reluctance because Jones had a "rougher" voice and he was two inches shorter than her, which was much shorter than the previous height descriptions she had given of her assailant. She told the jury she was "98 percent sure" of her identification. On February 6, 1973, Wilbert Jones was sentenced to life in prison. In March, 2015, the Innocence Project in New Orleans reviewed Jones' case and found that the prosecution had not disclosed evidence that another man with gapped teeth committed similar attacks shortly after the nurse was assaulted. In October, 2017, Jones was granted a new trial and, a year later, the prosecution dismissed the charges against him. Jones was released from prison after serving 45 years of a life sentence for a crime that he did not commit (Possley, 2018).

In this instance, the guilty culprit was already serving prison time for other crimes, however, too often the true culprit is never found and an innocent person is arrested and incarcerated. Similarly, in the United Kingdom, Irish postman, Victor Nealon, was accused and convicted of attempted rape despite eyewitnesses describing

a man that differed from Nealon's physical appearance and selected another individual in the lineup parade. Testimony revolved around the perpetrator's accent that was presumed to be Scottish. Nealon, who had a distinctive Northern Irish accent, was at home with his girlfriend during the incident. Nevertheless, he was arrested and tried for the attempted rape. Despite a lack of forensic evidence, he was convicted and served 17 years of a life sentence before being exonerated by DNA evidence (R v Nealon, 2014). For the fortunate few, persistence and DNA evidence led to successful appeals. However, in cases like that of Wilbert Jones, where the evidence was destroyed, there is little chance or hope for exoneration.

1.2 Eyewitness Identification is Flawed

“A witness' voice memory is not exempt from the sort of problems that we more commonly associate with a witness' vision; just as with eyewitness identification, expert testimony on the reliability of voice identification reveals vulnerabilities that lie outside the range of common knowledge.” (Schiro, 679 F.3d at 534, as cited in Saltzburg, 2013).

In the US, more than 70% of wrongful convictions were primarily due to inaccurate eyewitness identification (The Innocence Project, 2014). At present, both eyewitness and earwitness testimony are permitted in court trials. The question remains as to how much reliance can be placed on the accuracy of this information and whether it should be regarded as reliable evidence in police lineups and subsequent court testimony (Goldstein, Chance, & Schneller, 1989) Typically, eyewitness testimony holds significant weight and credibility when presented as evidence during trial. Jurors are more likely to associate a witness' high identification confidence with accuracy (Semmler, Brewer, & Douglass, 2012). Although

eyewitnesses may struggle during the initial identification process, admittance or allusion to identification uncertainty is not introduced during trial, leading legal counsel and jurors to give substantial credence to the evidence. Most testimony is presented after a lengthy amount of time has passed since the occurrence of the crime and the subsequent identification (Deffenbacher et al., 1989).

Several wrongful convictions in the last few decades have isolated eyewitness identification inaccuracies as one of the most crucial flaws in the justice system (Loftus, 2005; Penrod & Cutler, 1995). Stories like Nealon's and Jones' are still prevalent, and more research must be conducted on eyewitness and earwitness identification to reduce misidentification and wrongful convictions. Within the context of eyewitness inaccuracies, earwitness misidentification accounted for 17 out of over 350 wrongful convictions in the United States and was vital prosecutorial evidence in five trials (Sherrin, 2015). Although there are no comprehensive statistics to account for convictions based on eyewitness identification in the UK, well-known convictions like Victor Nealon's shed light on these injustices. Unlike the extensive research on eyewitness identification, earwitness identification is less studied and new strategies are uncovered everyday by researchers to enhance identification accuracy or, at a minimum, to reduce misidentification of the wrong perpetrator.

1.3 What is Earwitness Identification?

Earwitness identification is the process of a witness hearing the voice of a target person or persons, retaining that information in memory, retrieving that information later when called to identify the suspect(s) either in a 1-person voice lineup or a [multiple]-person voice lineup, and finally, testifying or communicating this decision to a police investigator, trial judge, or jury (Yarmey, 1995, p.795).

Many researchers have investigated the accuracy of eyewitness identification and testimony; however, there has been less of a focus on earwitness identification, although it has been used as evidence in lineups and trial testimony. Although earwitness identification is typically investigated under the broader framework of eyewitness identification, it stands apart from the traditional recognition processes associated with eyewitness identification. There are mixed reviews supporting the use of eyewitness identification tactics as a basis for speaker identification. Hollien (2012) noted that eyewitness and earwitness identification vary in relation to the processes of visual and auditory memories, the anatomical structure of the eyes and ears, how emotional states may affect the identification process, how vision impairments contrast with hearing impairments, and the innate ability of individuals to remember visual or auditory input. These fundamental differences suggest that earwitness identification accuracy should be reviewed specific to the auditory processes of speaker identification rather than the visual processing methods used in eyewitness identification (Hollien, 2012). Approaches for eyewitness identification are not analogous to earwitness identification and new strategies that acknowledge the relevant sensory input mode should be applied.

The paradigms of eyewitness identification, face recognition, and earwitness identification may vary in the number of targets presented, the type and duration of exposure to the targets, and the format of the recognition task that follows the exposure (Lindsay, Mansour, Bertrand, Kalmet, & Melsom, 2011). While eyewitness identification paradigms may have been conceptualized using the theories from facial recognition paradigms, the methodologies used by researchers tend to differ. Eyewitness memory researchers tend to present a single target face within a diverse background at various exposure durations. Following exposure and encoding,

witnesses are presented with a target-present or target-absent lineup in either a live or simulated format rather than using photographic stimuli (Lindsay et al., 2011). Conversely, face recognition researchers tend to present several targets on a bare background without distinguishable facial features. There is no emphasis on encoding as witnesses are privy to the recognition task prior to exposure and are usually presented with the same photos from the study task (Lindsay et al., 2011).

Earwitness identification employs similar patterns from both eyewitness memory and face recognition research. Targets may be presented either in a single or multiple presentation format (Smith et al., 2020). There is a range of exposure durations and the witnesses are unaware of a recognition task. Earwitnesses are presented with a target-present or target-absent voice lineup that may include the same stimuli or a variation of the stimuli presented during the study task. For eyewitness and earwitness identification, the objectives of the ongoing research are to generate new identification strategies for fair lineup procedures that will impact policy decisions (Wells, 2001).

1.4 Earwitness Identification in History

Earwitness identification has been documented as early as the 1600s. It was most notably used as critical evidence in the Charles Lindbergh kidnapping case, in which Bruno Richard Hauptmann was on trial for kidnapping and murdering Lindbergh's son. Lindbergh testified that Hauptmann's voice was the same voice he heard say, "Hey, Doc, over here" three years earlier during a ransom drop, although he said that it would be difficult to identify Bruno just by his voice (*State v. Hauptmann*, 1935). Based on this testimony and circumstantial evidence, Hauptmann was sentenced to death and executed.

Compelled by the earwitness evidence presented in the Lindbergh case,

McGehee (1937) investigated the reliability of voice identification evidence. She examined factors that affected voice identification and primarily focused on how accurately men and women identified voices they had previously heard after various durations of time, known as retention intervals. According to McGehee, as the interval of time increased between the first encounter of the voice and the subsequent identification, the identification accuracy decreased from 83% after one or two days to 13% after five months. In many cases, witnesses and victims are asked to identify a perpetrator's voice after a lengthy delay from the initial exposure to the identification (Kerstholt, Jansen, Van Amersvoort, & Broeders, 2004). The length of time between their exposure to the speaker's voice to the actual identification lineup can be weeks, months, or years. Additionally, the duration of time an earwitness is exposed to a speaker's voice is critical. In the Hauptmann case, Lindbergh heard the perpetrator speak a few words (*State v. Hauptmann*, 1935) which impacted his exposure to the speaker's voice. Although research has suggested that at least a 2-second exposure duration can produce an accurate identification (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012), longer exposure times will increase accuracy (Cook & Wilding, 1997b). This thesis will further examine how the length of exposure time and retention intervals may impact earwitness memory processing and contribute to poor performance in speaker identification.

1.5 Who is an Earwitness?

Although linguistic and forensic experts are called to offer expertise on earwitness evidence presented during court trials, most earwitnesses do not possess exceptional expertise in speaker identification (Robson, 2018). Earwitnesses are layperson listeners (Yarmey, 2012) with the innate ability to identify acoustic sounds and voices they experience every day (Nolan, 1997). They can be anyone in the

general population who possess specific qualities and physical traits that may help or hinder their ability to identify voices.

For layperson listeners, acoustical voice characteristics present memory processing challenges beyond the typical visual perspective. Variables in pitch, tone, speaker rate, and other aspects make earwitness identification exceptionally problematic; other factors, including age, gender, and mode of voice presentation at the time of the event, adds further complexity (Mullennix et al., 2010). Multiple factors that impact the accuracy of identification evidence and relying heavily on that evidence to apprehend perpetrators and support court testimony could have detrimental consequences. This thesis will examine how an earwitness' individual attributes like native language and accents and physical traits like age and gender may impact how well an earwitness remembers voices. I will also explore how aural characteristics unique to the speaker like familiarity, pitch, tone, speaking rate, and distinctiveness can influence an earwitness' memory for the speaker's voice and impact speaker identification accuracy.

1.6 Confidence

Witnesses and victims are restricted to providing identification evidence based solely on aural exposure when they are unable to view the perpetrator. In such circumstances, witnesses and victims may be convinced that they can remember the perpetrator's voice and, often, support the auditory evidence with high confidence (Yarmey, 2012). Remembering a speaker's voice while experiencing an emotional or traumatic event can be quite difficult for earwitnesses. Witnesses may exude confidence that they would never forget a particular voice and offer that confidence after making an identification. This level of confidence is presented as evidence in subsequent court testimony, and too much value is placed on the identification (Howe,

Knott, & Conway, 2017, p. 61) while discounting the errors that have and will occur. This thesis will examine the relationship between confidence and identification accuracy and discuss the implications of providing confidence ratings to earwitness evidence.

1.7 Earwitness Memory

Earwitness memory is the recall and recognition of auditory information by witnesses (Heath & Moore, 2011). Crimes committed by perpetrators in disguise, by telephone, or in poor viewing conditions make accurate visual identification difficult and reliance on auditory information essential. In situations where the witness did not see the perpetrator, they can only identify the individual based on his/her voice. After witnessing a crime, it is easy to forget details like height, facial features, or voice characteristics. In Jones' case, the nurse changed the height of her assailant three times and told police that he had gapped teeth and a soft voice. The victim chose Jones in a lineup despite his "rougher" voice and shorter stature. Eyewitnesses to a significant event like a crime believe they will remember a face or a voice and will be able to identify the perpetrator in a lineup (Sherrin, 2015). Memory interference from post-event information or verbal overshadowing may alter the earwitness' initial memories of the perpetrator and lead to identification errors.

Research on the inner memory processes involved in earwitness identification has helped to shed some light on what occurs when we encode and later recall a witnessed event. Human memory consists of three basic processes: encoding, storage, and retrieval (Melton, 1963). Encoding, storing, and retrieving the information required to later recognize a voice in a lineup is critical. These processes form the basis of memory, but more complex memory systems determine which memories are imprinted and later retrieved when it is necessary to recognize or recall part of the

stored information. The process of encoding entails transforming sensory information into a type of input that can be stored in the memory (Nevid, 2013). Typically, the witness observes the event and encodes it in their memory for storage and, later, retrieval. Tulving and Thomson (1973) proposed the *encoding specificity principle*, which stipulates that the retrieval of stored information is best when the retrieval cues share the same or similar stimuli that were present when the information was encoded. After encoding, we store or retain, the information in our memory and, later, we call on that information either to recognize something familiar or to recall something more extensive (Jacoby, 2010). Often, it is essential to remember the speaker's voice and the words or phrases they spoke during the crime. The present thesis will discuss the memory systems involved in retaining memories for voice characteristics and information content. I will discuss how memory interference occurs and how it impacts identification accuracy.

1.8 Forensic Implications

During the Hauptmann trial, the complications of speaker identification were not addressed at the time because it was understood that people had an innate ability to recognize voices (Nolan, 1997). After the Hauptmann trial, researchers took notice of earwitness identification and began to investigate it from a psychological perspective. If this trial were to take place today, more questions surrounding Lindberg's identification would likely need to be answered in order for his earwitness evidence and testimony to be admissible. The present thesis will discuss earwitness identification from a psychological perspective and evaluate the current voice identification lineup procedures in the U.S. and the UK. Lastly, I will discuss new technological developments implemented by researchers and law enforcement that may present an effective strategy to reduce misidentification in the future.

1.9 Chapter Summaries

Chapter 1 introduced earwitness identification as it applies in the context of the broader research of eyewitness identification. We identified that witnesses, or listeners, do not possess superior expertise in speaker identification and tend to be witnesses out of happenstance. As earwitnesses, strategies that support identification accuracy should differ from the strategies used in visual eyewitness identification. When earwitnesses attempt to identify unfamiliar speaking voices, many aural characteristics impact the identification process that may reduce accuracy. We addressed issues during the encoding process that may hinder the retention of a speaker's voice and subsequently impact identification for testimony purposes.

Chapter 2 highlights the factors that influence speaker identification and impact voice lineup procedures. Our research focuses on unfamiliar voices because voice lineups will be futile when identifying familiar voices. Variability of aural characteristics of a speaker's voice like pitch and distinctiveness may influence a witness's ability to make a correct identification. A listener's age may attribute to an increase or decrease in identification accuracy and an own-gender bias for voices of the same gender may lead to more accurate identifications than opposite gender voices. Verbal overshadowing and source confusion may occur before providing identification evidence that may skew accuracy and there are mixed results as to the reliability of confidence ratings in relation to identification accuracy. Exposure length and retention intervals can aid in the encoding process for later memory retrieval and the memory processes involved in speaker identification were further addressed. The chapter also explores the forensic implications of lineups in the UK and U.S. and explains the recommended guidelines currently in place to maintain consistency across law enforcement departments.

Chapter 3 addresses how well people can recognize words and voices they previously heard once. In Experiment 1, we analyzed memory for monosyllabic words and memory for speaker identification for male and female voices. We predicted that words would be more accurately recognized than voices and female voices would be identified at a higher rate than male voices. We piloted new presentation formats in Pilots 1 and 2, and changed material content in Experiment 2 with a physically violent robbery scenario to analyze the participants' identification performance based on the gender of the robber and speaking voice. The justification for the provocative content is addressed and the results of the experiments and pilots studies are analyzed and discussed.

Chapter 4 examined the participants' ability to identify previously heard speakers and neutral interview content in Experiment 3. A further analysis to determine the effects of exposure duration was conducted on data from Experiments 1 and 3. We predicted that longer voice sample exposures would improved identification performance for the speakers and content material. Suggestions for extending the voice sample duration are addressed and the results for Experiment 3 and the cross-analysis of Experiments 1 and 3 are presented and discussed.

Chapter 5 investigates the type of content and presentation modality. In Experiment 4, provocative content was presented in a written or auditory form to determine the participants' ability to discriminate between original and altered statements. An assessment of the effect of the speaker's gender in the auditory format was conducted. We predicted higher accuracy in auditory statements and better performance in female voices. The justification for written and auditory modalities is explored and the results are discussed.

Chapter 6 addresses the impact of familiarity on speaker identification and

also assesses the relationship between identification accuracy and confidence ratings in Experiment 5. We predicted participants would more accurately identify familiar voices and confidence ratings would be higher for correct responses. A further exploration of the confidence-accuracy relationship and its utility in voice lineups was addressed.

Chapter 7 investigates the effects of provocative content and the gender of the crime victim in recall accuracy. Participants reviewed details of a crime and their recall accuracy was calculated from a simulated crime report. We predicted that recall accuracy would be higher when the victims are female and female participants would recall more details than their male counterparts. The context of the writing superiority effect is addressed and the results of the recall analysis are discussed.

Chapter 8 summarizes the findings of the previous chapters and discusses the limitations presented in the six experiments. Suggestions for future research in improving voice lineups and implementing digital modalities are discussed.

Table 1-1

Overview of the six experiments presented in this thesis

	Learning Phase		Testing Phase		
	Exposure	Retention Interval	Voice sample	Manipulated Variables	Dependent Variables
Experiment 1 (Word)	Heard a voice for 2s	Immediate*	40 voices	Old vs. New Male vs. Female	Memory for content
Experiment 1 (Voice)	Heard a voice for 2s	Immediate*	40 voices	Old vs. New Male vs. Female	Voice identification
Pilot 1	Heard a voice for 8s	Immediate*	40 voices	Old vs. New Male vs. Female	Voice identification
Pilot 2	Heard a voice for 4s/8s	Immediate*	Attended to previous voice sample	Male vs. Female	Voice identification

Experiment 2	Heard a voice for 2s	Immediate*	40 voices	Old vs. New Presentation Format (Attendant vs. Robber)	Voice identification
Experiment 3	Heard a voice for up to 30s	45 seconds	30 voices	Male vs. Female Old vs. New	Voice identification Memory for content
Experiment 4 (written)	Read crime scenario	45 seconds	20 written scenarios	Old vs. Altered	Memory for content
Experiment 4 (audio)	Heard a voice for up to 30s	45 seconds	18 audio scenarios	Old vs. Altered	Memory for content
Experiment 5	Heard a voice for up to 30s	10 minutes	20 voices	Familiarity Old vs. New Confidence rating	Voice description Voice identification
Experiment 6	Heard a voice for up to 30s	10 minutes	1 voice	Crime Type Male vs. Female Recall	Memory for content

Note: *Under 10 seconds from exposure to test

Chapter 2 – Literature Review

Extensive research within the area of eyewitness identification falls short of an in-depth investigation into earwitness evidence. In cases like Wilber Jones, who was wrongfully convicted of rape primarily due to earwitness testimony, we can surmise that earwitness identification is flawed (Heath & Moore, 2011). Limited research in this area has failed to find strategies that may significantly reduce earwitness identification errors successfully. Earwitness identification stands apart from eyewitness identification due to the nature of acoustical characteristics that are involved as well as other variables like age and gender. Earwitness identifications center around a lay witness who is exposed to a speaker's voice (Stevenage, Howland, & Tippelt, 2011). Typically, this is likely to occur in crimes committed either when the perpetrator is in a disguise or cases of harassment over the phone (Yarmey, Yarmey, Yarmey, & Parliament, 2001).

After the witness experiences the event, they may be asked by law enforcement to identify the perpetrator in a voice lineup. When witnesses are requested to identify a perpetrator, the identification process involves a voice parade or lineup. In the lineups, voice samples are presented to the witness in a serial or sequential presentation. In a serial lineup, several voices are played in sequence and the witness provides a response after hearing all of the voice samples. In a sequential lineup, one voice sample is played with the intention of gaining a response from the eyewitness after each voice sample has been provided until a correct identification is made. Identification errors that are likely to occur during the lineup process are the focus of this thesis.

A lack of exploration has left earwitness research in a state of uncertainty that does not bode well for future witnesses, law enforcement, or the erroneously accused

suspects. Further psychological evaluation of earwitness identification is warranted and needed because ongoing research may cultivate better approaches to identification efforts and prove useful to law enforcement officials and legal professionals. This chapter examines earwitness identification and addresses how earwitnesses perform in identification tasks. I will explore how earwitnesses are vulnerable to factors that impact identification accuracy, discuss the memory systems involved, and explain how those processes affect an earwitness's ability to identify a perpetrator successfully.

2.1 Factors that Affect an Earwitness' Identification Performance

Familiarity

Earwitness research has explored exposure to familiar and unfamiliar speakers. As we have previously mentioned, witnesses tend to identify familiar speakers better than unfamiliar speakers (Yarmey, 2012). However, when voices of familiar speakers are disguised, speaker identification accuracy declines (Yarmey et al., 2001). Read and Craik (1995) found that listeners performed poorly in speaker identification despite expressing a strong familiarity with the speaker. Yarmey et al. (2001) examined various familiarity levels of high, moderate, low, and unfamiliar ratings. They found that participants identified high and moderately familiar voices (85% and 79%, respectively) much better than low and unfamiliar voices (49% and 55%, respectively). They also found a higher rate of false alarms for the low (23%) and unfamiliar (45%) voices than the high (5%) and moderately (13%) familiar voices. Conflicting results on identifying familiar voices persist and difficulties can exist in identifying voices of family members, native language speakers, regional dialects, and accents.

Native and non-native language speakers may differ in identification accuracy based on their familiarity with the speaker's language (Philippon, Cherryman, Bull, & Vrij, 2007). Native English speakers were presented with a video of a person speaking to an accomplice either in English or French. All facial traits were hidden to only expose the listeners to the speaker's voice. Listeners were asked to select the voice from either a target-present or target-absent lineup. The listeners correctly selected the speaking voices in English and French equally well (46.7% for English and French) in the target-present condition; however, there were significantly more false alarms for the French-speaking voice than the English voice (46.7% and 20%, respectively). In the target-absent condition, listeners correctly rejected more English-speaking voices (33.3%) than French voices (6.7%) and false alarms were very high for both speaking voices (66.7% for English voices and 93.3% for French voices).

Familiarity in speaker identification transcends a prior knowledge of a person's speaking voice. Aural characteristics like regional dialect and accents can also present complications in speaker identification. In some instances, witnesses may be requested to identify voices spoken with a national or regional accent. This presents an additional obstacle in identification because witnesses may find unfamiliar accents more challenging to identify (Pickel & Staller, 2012). Accents can vary among speakers of foreign languages and occupants of specific regions within a populous. Given the diverse demographics within large urban cities, it is likely that daily encounters with non-Native language speakers and accented speakers will occur. In eyewitness identification, witnesses visually identify people of their same race more accurately than others of a different race (Yarmey, 2012). Comparatively speaking, a similar effect may occur in speaker identification when accents are present.

In research involving Native and non-Native speakers, U.S. participants identified U.S. Native speakers more accurately (88%) than non-Native foreign speakers (13%), and English participants performed better with Native speakers (87%) than non-Native speakers (12%) (Doty, 1998). The “other-accent” effect also occurred among participants of the same nationality. In another study, Dutch participants listened to a voice lineup that included a regional accent from the Hague, which was considered a non-standard accent to the participants. In a target-present lineup, participants correctly identified only 24 percent of the target voices due to the impact of the non-standard regional accent (Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2006).

Stevenage, Clarke, and McNeill (2012) examined whether an “other-accent” effect would arise in a speaker identification lineup involving English and Scottish voices. Researchers presumed that Glaswegian witnesses would have been exposed to English accents more frequently than English witnesses to Glaswegian accents, thus creating a disproportionate effect that is commonly found in the “other-race” effect in eyewitnesses. In target-present lineups, English witnesses correctly identified speakers with an English accent better (73%) than those with a Glaswegian accent (53%), whereby Glaswegian witnesses performed only slightly better with the Glaswegian accent (58%) than the English accent (42%). False alarm rates were high for English witnesses (43%) and Glaswegian witnesses (42%). When applied in a real-life setting, the “other-accent” effect further suggests that earwitness identification is prone to error. Changes in aural characteristics, familiarity, and accents make accurate identification very difficult and unreliable. Introducing voice identification as evidence in court should be done with great caution and a detailed explanation of its suggestibility should be provided for jurors. Although the present thesis did not

analyze identification with respect to language or accent variations, all voice samples exhibited a neutral, South East England accent or a neutral London accent. Each participant was exposed to or had familiarity with all voice sample accents. In Experiment 5, participants rated their familiarity with the voice samples on a Likert scale to provide insight on whether familiar voices are better recognized than unfamiliar voices. The other five experiments (and two pilot studies) did not provide a familiarity rating option and presumed that the presented voice samples were unfamiliar to the witness to demonstrate how witnesses perform when identifying unfamiliar voices.

Gender

One notable distinction in speaker identification is gender, that is, the speaker's gender and the witness's gender. A speaker's voice is usually categorized as a male or female based on certain acoustical attributes (Pernet & Belin, 2012). The challenge researchers face is how well witnesses recognize male and female voices (Campeanu, Craik, & Alain, 2015). Previous research findings have indicated a potential gender-bias whereby males recognize male voices better than female voices and vice versa (Roebuck & Wilding, 1993). Initial research by McGehee (1937) showed that male witnesses were generally better able to identify voices than female witnesses. Males more accurately identified male voices than female voices indicating that there may be a gender bias among male witnesses.

Conversely, a comprehensive analysis of five studies by Cook (as cited in Wilding & Cook, 2000) showed a significant interaction between the gender of the witness and the gender of the speaker. In each experiment, witnesses heard one unfamiliar male voice and one unfamiliar female voice speak a sentence. They were later presented with a lineup of six male voices and six female voices and had to

identify the target voice for each lineup. Overall, female witnesses were more accurate in identifying female voices (51%) than male witnesses (38%). However, females were only two percent better at identifying male voices than male witnesses were (43% to 41%, respectively). Comparably, Aglieri et al. (2017) found that male and female listeners identified females voices better than male voices and females listeners showed a gender-bias for female voices, whereas males did not. Overall, there was not a significant effect of the listener's gender on speaker identification.

Not all research supports a female speaker identification bias for female witnesses, especially when familiar voices are involved. When German-speaking male and female students identified the voices of fellow male and female classmates, there were significant main effects of voice gender and witness gender (Skuk & Schweinberger, 2013). Overall, female witnesses identified familiar voices more accurately than male witnesses and male voices were identified at a higher rate than female voices by both genders. There was a slight gender-bias for male witnesses who identified male voices more accurately than female voices (39.5% and 27.9%, respectively). However, female witnesses identified both male and female voices with nearly the same accuracy (39.3% and 42.8%, respectively). Moreover, female witnesses identified male and female voices better than their male counterparts when they had engaged in frequent contact with the speaker.

In England and Wales, women make up 4.5% of the total prison population of 83,665 as of November 2019 (Prisonstudies.org, 2019a). In the U.S., women make up 9.8% of the total prison population of 2,121,600 as of December 2016 (Prisonstudies.org, 2019b). There is a large disparity in the gender of prison populations and, given the statistics, it is more than likely that a witness will encounter a male perpetrator much more often than a female perpetrator. Differences in

recognition related to both witness and speaker gender are evident. It is difficult to definitively determine a gender bias because the accuracy rates for male and female witnesses tend to vary. Research findings indicate that female witnesses may perform better overall than male witnesses, but that performance level differs when familiar and unfamiliar voices are involved (Skuk & Schweinberger, 2013). The present thesis extends the research on gender differences and examines the witnesses' performance in identifying male and female speakers.

Age

Memory variations can occur with age. Previous findings show that children have more difficulty identifying unfamiliar voices than younger adolescents (Yarmey, 2012). As we age, our ability to process auditory information changes. Adults tend to apply specific markers to a speaker's voice and later rely on them when exposed to a newly heard voice (Namy, Nygaard, & Sauerteig, 2002). The marker placement identifies distinctive characteristics like pitch, tone, and speaking rate and assigns those unique characteristics to a specific speaker. When the witness hears similar voice characteristics, they determine if it is a voice that is familiar or unfamiliar (Yarmey, 2012).

Children tend to recognize familiar voices better than unfamiliar voices and identification ability improves as they progress to adolescence (Yarmey, 2012). Both children and adults face challenges with speaker identification accuracy. Young adults up to the age of 40, outperform children and elderly witnesses in identification accuracy (Yarmey, 2012). Öhman, Eriksson, and Granhag (2011) investigated differences in unfamiliar speaker identification among children 7- 9 years old ($M = 7.96$, $SD = 0.54$), 11-13 years old ($M = 12.54$, $SD = 0.57$), and adults ($M = 30.26$, $SD = 10.97$). All three groups listened to a 40-second voice sample of a simulated phone

conversation planning a crime with non-distinctive (neutral) content. After a two-week retention interval, participants took part in either a target-present or target-absent lineup to identify the speaker. Participants were instructed that the voice may or may not be in the lineup. They were presented with seven full-length voice samples and instructed to listen to all the samples once and then listened to a shorter version of each sample a second time. They listened to all the samples or stopped after they heard to correct voice. They could also indicate that the voice was not present. They were asked to rate how sure they were of their choice.

Results for the target-absent lineup showed that all three groups performed above chance level (25%); however, for the target-present lineup, the second youngest group (27%) outperformed the youngest group (14%) and the adults (20%). Confidence ratings did not support a significant relationship between confidence and accuracy. The adults had the highest number of false alarms out of all the groups in both target-absent (60%) and target-present (50%) lineups compared to the youngest group (49% and 36%, respectively) and the second youngest group (49% and 46%, respectively). The results for the youngest age group supports research that younger children struggle with identification accuracy but contradicts findings that adults perform better in voice identification than children (Yarmey, 2012). Although the results conflict with previous findings, the overall results confirmed that witnesses do not perform well when identifying unfamiliar voices.

In criminal conversation cases, sometimes what was said determines how well it is remembered (Öhman, Eriksson, & Granhag, 2013). For example, conversations involving obscene content heard over the telephone were more accurately remembered by adults than children. The children remembered the content of the message, but refrained from reporting the sexual content (Leander, Granhag, &

Christianson, 2005). In adult witnesses, it is likely that stimulating conversation content like sexual or violent details are recalled more than neutral conversation content (Pezdek & Prull, 1993).

Exposure Time

Another factor that should be addressed when examining earwitness identification is the impact of the listener's exposure to a speaker's voice and how long they were exposed to the voice. In the Hauptmann case, Charles Lindbergh was exposed to just a few words spoken by the perpetrator that amounts to a very short duration of exposure time. Previous research suggests that short exposure times of two seconds is sufficient to make an accurate identification (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012); however, longer exposure times tend to make a better impact in not only increasing accuracy rates but also reducing false alarms (Kerstholt et al., 2004). Cook and Wilding (1997) found that when presented with short and long utterances, witnesses correctly identified more voices in the long-utterance condition than the short utterance condition.

Repeated or multiple exposures can also increase accuracy rates as it lengthens the exposure time of hearing a speaker's voice (Deffenbacher et al., 1989). Repeated exposure after a delay of two weeks showed that witnesses were able to identify voices better than the witnesses exposed to the voice for a continuous duration of time; however, performance overall was low (Deffenbacher et al., 1989). In the legal context, exposure time is critical in earwitness identification, especially when applying guidelines that account for suggestibility in the evidence. This thesis will address earwitness evidence as it applies to voice lineup procedures in the U.S. and the UK, later in this chapter.

Retention Interval

The time that lapses from the initial exposure of a voice to the time that the witness may be approached for a lineup is called the retention interval (Kerstholt et al., 2004). In real-life experiences, witnesses may be delayed from making an identification for weeks, months, or even years after witnessing a crime. Previous research has shown that the retention interval impacts how well witnesses perform on an identification task. Research has shown mixed results for accuracy. Some researchers suggest a decline in accuracy for delays up to 3 weeks (Yarmey & Matthys, 1992), and other researchers do not show any effect of retention delays on accuracy (Kerstholt et al., 2006). Although retention intervals were not a factor within the context of experimental analyses in this thesis, it is discussed as a relevant influence in identification accuracy.

Confidence

When presented with a voice sample for identification, witnesses may be asked to rate their confidence level of the identification. Identifications involving very familiar voices have shown a stronger relationship between accuracy and confidence than unfamiliar voices (Yarmey et al., 2001). Similarly, witnesses who correctly identified sentence length utterances rated their confidence to be much higher for those identifications than the incorrect identifications (Skuk & Schweinberger, 2013). However, previous research has shown mixed findings when analyzing accuracy and confidence ratings, as some researchers have not found a correlational relationship between confidence and accuracy (Öhman, Eriksson, & Granhag, 2011). Witnesses were not more accurate in identifying voices when they gave a higher confidence rating for their identification (Read & Craik, 1995) nor did providing positive feedback to witnesses significantly influence confidence ratings on memory recall (Rechdan et al., 2017). The relationship between identification accuracy and

confidence ratings is not definitive and may border on suggestibility if consistently relied on as valid evidence. It can be misleading to assign substantial evidentiary value to speaker identifications that are presented with higher confidence ratings. This thesis provides an additional contribution to the issue of confidence by examining the relationship between identification accuracy and confidence as it relates to voice familiarity.

Aural characteristics

Although the experiment analyses in this thesis did not apply aural changes, the following discussion on aural characteristics highlights the specific variations that may impact identification accuracy. Further examination of linguistic variances is beyond the scope of this thesis.

Pitch

In earwitness identification, aural characteristics can significantly impact the accuracy of an identification. In situations where voices may have been initially disguised by changing the pitch or inflection like whispering, witnesses struggle to correctly identify voices in subsequent lineups (Zetterholm, Sarwar, Thorvaldsson, & Allwood, 2012).

Pitch is defined as the “perceptual correlate of fundamental frequency (F0),” which is quantified in Hertz (Hz) (Spence, Arciuli & Villar, 2012). Pitch is one of several “surface properties” of speech that also includes tone, speaking rate, and amplitude (Laver, 1968, as cited in Mullennix et al., 2010). Pitch for adult men and women have a standardized range of 100 – 150 Hz and 175 – 250 Hz, respectively. Based on these specifications, men typically speak in a low pitch voice and women in a high pitch voice. Changes in pitch from the initial exposure of a voice can make accurate identification very difficult (McGorrery & McMahon, 2017). McGehee

(1944) investigated several characteristics that impact speaker identification accuracy, such as pitch, agreeableness, and speech rate. She discovered that voices with a low pitch and slow rate of speech were most likely to be misidentified unlike the other voices presented. Research has shown that the sensitivity to variations in pitch can impact identification (Mullennix et al., 2010). Other characteristics like the rate of speech, tone, accents, and distinctiveness present a way of disguising one's voice and create an obstacle for accurate identification.

Rate of Speech

Similar to pitch, rate of speech adds another layer of difficulty to accurate identification. The rate of speech varies from slow, moderate, or fast levels and is analyzed in units per second in words, syllables, and phonemes. Speaking rate is not always distorted as easily as other characteristics like pitch. Mullennix et al. (2010) explored whether a witness identified a voice as speaking slower or faster than the target voice rate based on previously stored information about the speaking rate of the voice. This speaking rate memory bias can be linked to difficulty in the encoding process. Upon further analysis of speaking rate and pitch, Mullennix et al. (2010) discovered that, unlike voice pitch, there was no memory bias for speaking rate. Surface properties of the voice like pitch and tone are encoded differently than characteristics like amplitude. Surface properties are automatically encoded and preserved in the memory, along with the material context of the speaker's message (Laver, 1968, as cited in Mullenix et al., 2010).

Variability in speaking rate from the initial exposure to the identification process has led to issues with encoding and subsequent errors in identification. Bradlow, Nygaard, and Pisoni (1999) analyzed talker variability, speaking rate variability, and amplitude. They found that a listener's spoken word recognition

accuracy decreased when they were presented with a different talker rather than the same talker across both short and long lag times. Short lag time was analogous to short term processing, and long lag time represented long-term memory retention. Similarly, listeners performed better with words spoken at the same speaking rates across short and long lag times. The findings indicated that words spoken at the same speaking rate were encoded and retained in the long-term memory rather than words spoken at different speaking rates. No difference was found for amplitude.

In a real-life context, a perpetrator may speak quickly during the commission of a crime and, later, speak at a slower rate when using a natural voice during the identification process. Research has shown that witnesses identify voices spoken at the same rate during the initial exposure and the identification process much better than voices spoken at different rates (Bradlow et al., 1999). Earwitnesses exposed to a speaking voice rate that differs from the encoded voice rate may find it challenging to identify the correct perpetrator.

Tone

Comparable to speaking rate variability, tonal differences are retained in the long-term memory. Tonal changes from emotional influences such as anger or fear, whispering, or deliberate disguise can impact speaker identification accuracy. An altered tone of voice that is different from the initial presentation reduces identification accuracy (Saslove & Yarmey, 1980). This is critical in application because most voice parades are re-recorded statements or pieces of conversational dialogue where changes in tonal characteristics have occurred. In one study, when witnesses heard an emotional phrase, their ability to accurately identify the same voice 17 days later when it was presented in a low to moderate tone decreased to 17%. However, accuracy increased to 66% when the voice was presented in the same

speaking voice and emotional tone (Read & Craik, 1995). Subsequent results showed that re-recording the same voice with similar emotionality also increased accuracy but not at the same rate as presenting the identical emotional voice. Higher confidence levels were also associated with the presentation of the same emotional tones compared to the re-recorded and low to moderate tones. These results demonstrate the significant consequences that tonal variations can have on earwitness identification.

In real-life criminal encounters, witnesses are inadvertent observers under substandard conditions, which make the subsequent task of identifying voices difficult. In ideal testing conditions, identification accuracy remains no better than chance when the recognized voice is the same tone as the initial voice (Read & Craik, 1995). Participants who heard an angry voice followed by a voice spoken in a normal tone, correctly identified the voice at just above chance level (16%, where chance is 12.5% based on eight possible responses) when responding after a short delay compared to below chance (9%) when responding after a two-week delay (Öhman, Eriksson, & Granhag, 2013). Ideally, attempts should be made to match the voices' emotional tone in a lineup to the initial voice heard by the witness. Although duplication of an exact tonal match is impossible, a near similar tone presented during the lineup process will likely improve identification accuracy.

Distinctiveness

Similar to pitch, speaking rate, and tonal characteristics, a voice's distinctive features can affect identification accuracy. The issue with determining distinctiveness is that the definition is subjective to each researcher. Orchard and Yarmey (1995) distinguished between distinctive and non-distinctive voices based on the higher ratings assigned to voices that were “highly striking” by participants. They evaluated the identification accuracy rates of whispered and normal speaking levels in both

distinctive and non-distinctive voices. Hit rates were higher in the whisper-whisper and normal-normal conditions for distinctive voices than non-distinctive voices within target-present lineups. However, correct rejection rates were much higher in the normal-normal voices in non-distinctive voices compared to all the other conditions. Conversely, the reduction of distinctive features when whispering, for example, will reduce accuracy rates. Participants were less likely to correctly identify the same whispered speaking voice (18%) of an unfamiliar speaker than a highly familiar speaker (35%) (Yarmey et al., 2001). This discrepancy necessitates a more consistent definition of distinctiveness to determine how it may impact speaker identification.

Super Recognizers

Super recognizers are known to have an exceptional ability to recognize unfamiliar faces with a high level of accuracy above the average layperson (Russell, Duchaine, & Nakayama, 2009). Super recognition ability is not only valuable for visual perception but also auditory recognition. Super voice recognizers have the ability to remember and recognize voices extremely well (Aglieri et al., 2017). On the other end of the spectrum, individuals diagnosed with developmental phonagnosia are unable to recognize voices including familiar voices like celebrities and family.

To test the performance of speaker identification and identify super voice recognizers, Aglieri et al. (2017) created the Glasgow Voice Memory Test. They investigated how well people can remember unfamiliar voices as well as non-vocal stimuli (i.e., a bell). The participant group consisted of 1,120 lay listeners and one subject diagnosed with developmental phonagnosia. During the encoding phase, listeners heard voices and bell sounds. They later had to distinguish between old and new voices and old and new bell sounds. Listeners with a standard deviation of 2 or

higher above the percent correct score mean (i.e., the hit rate and correct rejections) were considered super voice recognizers (Roswandowitz, 2014, as cited in Aglieri et al., 2017). Conversely, anyone with a score of 2 standard deviations below the positively correct mean were considered potentially displaying signs of phonagnosia. The single phonagnostic subject had significantly lower scores for both speaker identification, as well as sound recognition for the bell, thus confirming the memory test detection scheme. Overall, they did not find any participants with scores reflecting that of a super recognizer. The present thesis does not examine possible super recognizers in the experimental context but it offers an insight on witnesses who may have a superior ability to identify voices that differs from layperson witnesses.

2.2 Factors that Impact a Listener's Memory

Memory Processes

The basic processes of memory are encoding, storage, and retrieval. These memory processes illustrate how information is processed through our memory system. When witnesses hear or observe a crime, they encode, store, and later retrieve certain aspects of that event. Encoding transforms incoming information into a code and moves that code into storage where it remains temporarily or until it is accessed, or retrieved, for use (Holt, 2019). Tulving and Thomson (1973) developed the *encoding specificity principle* which states that retrieval conditions should be the same as the conditions present during encoding. Unfortunately, problems can arise and contribute to identification inaccuracies because of difficulties with encoding, storage retention, or retrieval processes within our memory system.

The most widely adopted memory system model is the Multi-Store Model developed by Atkinson and Shiffrin (as cited in Baddeley, Papagno, & Vallar, 1988). This model suggests that sensory memory, short-term memory, and long-term

memory work in sequence from input to retrieval. Sensory memory retains information that is relevant to the senses and allocates sensory codes for visual memory (i.e. iconic memory) and auditory memory (i.e. echoic memory) as well as the other senses (Holt, 2019). Earwitnesses utilize echoic memory to retain a certain amount of memory for auditory information for a short period of time (Read & Craik, 1995). Although echoic storage lasts longer than iconic storage, it can fall victim to decay and interference (Cowan, 1984). However, there is a possibility that the distinctive characteristics of a stored sound may reactivate auditory memory and allow the recognition of those characteristics to be retrieved (Winkler & Cowan, 2005). This is similar to creating voice markers to recognize a familiar voice that matches those markers (Yarmey, 1995).

Although information held in the sensory memory can decay, some information may be transferred into a code that will be stored in short-term memory (Holt, 2019). Short-term memory captures memory for a short period of time that will either fade or be retained in the long-term memory (Martin, Mullennix, Pisoni, & Summers, 1989). The short-term memory duration is quite short as the name suggests and its capacity to hold a large amount of information is limited. If the information held in the short-term is not rehearsed, the average duration of storage time is typically around 15-30 seconds (Atkinson & Shiffrin, 1971). After rehearsal, memories move from short-term memory to long-term memory.

Baddeley and Hitch (1974) expanded on the idea of short-term memory as a system where memory is a working process holding information before it is stored in the long-term memory. They classified this type of storage as a working memory. Working memory operates on information input for a duration of only a few seconds (Baddeley, 2003). Working memory tends to be categorized as short-term memory;

however, the two are quite different. Working memory involves both processing and storage, whereas short-term memory briefly stores an event. The *working memory model*, initially adopted by Baddeley and Hitch (1974), consisted of the central executive, the visuospatial sketchpad, and the phonological loop. The central executive controls the phonological loop and the visuospatial sketchpad processes. The phonological loop is comprised of the phonological store and articulatory rehearsal system, or loop, and stores the sounds heard either in spoken words or an internal voice. An earwitness utilizes the phonological loop when they hear an unfamiliar voice or repeats a name over and over. The rehearsal system is engaged when information is continuously repeated to retain it in the phonological store (Baddeley, 2003). The visuospatial sketchpad monitors visual images and spatial layouts. The sketchpad can function on its own or with the phonological loop. Evidence has shown that interference with either the visuospatial sketchpad or the phonological loop while learning a new task makes the task difficult to perform (Jaroslawska, Gathercole, & Holmes, 2018).

The central executive is the main component that oversees the other subsystems. The central executive acts as the lead operator that coordinates the other systems to operate in order to perform an action (Baddeley & Hitch, 1974). Baddeley (2000) later added the episodic buffer, which is a short-term storage space that moderates the interaction between all the systems to make stored information available for retrieval. It pieces all of the information together into a cohesive memory.

Long-term memory has a large storage capacity and can store information for a long duration of time. It consists of both implicit and explicit memory (Schneider, 2015). Implicit memory is a retrieved memory that lacks conscious awareness and

functions as procedural memory that exists in our actions and, at times, conditioned response. Explicit memory allows us to consciously recognize or recall a memory (Schneider, 2015). It functions as declarative memory which is factual knowledge that concerns our personal experiences stored in our episodic memory and concerns our knowledge for words and language stored in our semantic memory (Schneider, 2015).

Episodic memory allows a witness to recall or recognize important information from a crime, such as a date, location, time, and details about the people involved in the crime (Schneider, 2015). It is relevant to eyewitness identification because it relies on the witness to recall certain facts that occurred within an event or crime. Those facts can include relevant information leading to the apprehension of a perpetrator. While episodic memory may store details of words spoken during an event, semantic memory allows the witness to interpret meaning from those words or conversations observed during an event that may be beneficial in the identification process (Schneider, 2015). Like episodic memory, how the semantic memory is encoded and stored will impact the retrieval process.

In voice lineups, witnesses are asked to retrieve some information about a perpetrator so they make an accurate identification. When a perpetrator's voice is encoded in the long term memory, specific voice markers serve as retrieval cues that may later help to identify the correct voice (Palmeri, Goldinger, & Pisoni, 1993). The use of retrieval cues might help to elicit recall for a specific memory of the crime that may aid in a successful identification. For retrieval to be effective, the retrieval cues must be similar or match the cues present at encoding (Dewhurst & Knott, 2010; Tulving & Thomson, 1973). After witnessing a criminal event, the witness may be requested to not only identify a voice, but also remember what was said during the event. Memory for content rests on the witness's ability to encode the content to recall

or recognize that content later (Schneider, 2015).

Memory for Content

To understand how a witness can identify an unfamiliar voice that they have heard once and the content of the speaker's information, we must consider the memory processes at work during the presentation and identification phases. Often, after a perpetrator has been identified, the witness may be called to testify in court. Testimony may involve reporting the details of words or conversations that the witness heard. The conversation content may be critical for criminal investigations and court trials so the essential details are necessary to recall. The witness may feel compelled to retrieve the content verbatim, but that is nearly impossible (Neisser, 1981). Memory recall involves two particular systems that operate within the memory, gist memory and verbatim memory (Brainerd and Reyna, 1993). Gist memory is a small synopsis of a concept or phrase or remembering the peripheral context of that particular conversation. Verbatim memory is a complete detail of the conversation recalled as a word by word recollection (Brainerd and Reyna, 1993). Neisser (1981) defined verbatim recall as “word-for-word reproduction.”

Only in an ideal world can a verbatim memory of events exist. John Dean, the former White House Counsel to U.S. President Richard Nixon, was labeled “the human tape recorder” for his detailed testimony regarding the Watergate scandal (Neisser, 1981). He insisted that he did not remember the conversations he had with President Nixon verbatim but had a detailed recall of the events based on newspaper clippings he had saved. He provided a lengthy statement of these meetings only to discover later that the majority of the meetings held in the Oval Office were secretly tape-recorded. When the meetings were transcribed (by Pres. Nixon himself), the comparison of Dean’s testimony and the actual transcript proved to be quite different.

The overall acknowledgment and culpability of a criminal cover-up were evident, but the testimony did not even offer a gist of the conversations. Neisser (1981) suggested that Dean's recollections were "repisodic" in nature in that they were remembered solely because of the repetitive nature of specific phrases and reports rather than by verbatim recall. Recall differs from recognition because it requires an exact reproduction of a prior event that has been stored in the memory (Jacoby, 2010). By contrast, recognition involves a cue that triggers a recollection of a previous event. Generating memories with such detail as to recall the exact words of a conversation is extremely difficult. Even in the case of repeated exposure to an event or story, recall accuracy is limited, whereby recognition may be more accurate (Jacoby, 2010).

Face Overshadowing Effect

In eyewitness identification, there are occurrences when the witness has been exposed to the suspect's face, voice, or both. When witnesses are exposed only to a voice during the learning and testing phases, their performance accuracy is higher than when they identify a voice after being exposed to the face and the voice (Heath & Moore, 2011). This face overshadowing effect suggests that the strength of facial stimuli impacts how witnesses encode unfamiliar voices for subsequent recognition. The presence of a visual stimulus like a face distracts attention from the voice and leads to errors in speaker identification and impacts memory for content. This thesis focuses only on identification as it pertains to voices. Although audio-visual identification accuracy is not the focus of this thesis, it warrants mentioning because research in this area is limited. Most eyewitnesses will likely experience events that incorporate both visual and auditory modalities.

Verbal Overshadowing

Like the face overshadowing effect, verbal overshadowing is a challenge some

witnesses face when asked to describe the perpetrator's voice. By focusing on describing the voice, the ability to recognize the voice in a lineup is impaired (Schooler & Engstler-Schooler, 1990). In previous research, participants were asked to describe the facial features of a criminal suspect they witnessed committing a robbery. When they later had to select the suspect in a lineup, they struggled to find the target suspect. By verbalizing a description, the descriptive information overshadows the previous information that was encoded. Although this was initially tested in eyewitness identification, similar results have occurred when testing earwitness identification. When witnesses were asked to describe a voice before selecting the voice out of lineup, they struggled to identify the target voice (Perfect, Hunt, & Harris, 2002).

Verbal overshadowing has occurred when the initial encoded stimuli content has changed. Voice characteristics like pitch and speaking rate can vary within the same individual (Mullennix et al., 2010). In the time between the witnesses' exposure to the speaker's voice and identifying the speaker in a lineup, natural changes in the voice may have occurred (Zetterholm et al., 2012). Often during the commission of a crime, perpetrators will change or disguise their faces (Mansour et al., 2012) or their voices (Orchard & Yarmey, 1995) to make it difficult to identify them in a lineup. In stressful situations, a speaker may increase his speaking rate, for example, which will decrease when the speaker returns to a more relaxed state. Where the witness has heard a stimuli phrase during lineup that differed from the initial phrase spoken in the same voice, they struggled to identify the target voice when asked to give a description prior to lineup identification (Vanags, Carroll, & Perfect, 2005). The present thesis did not examine verbal overshadowing effects as it relates to earwitness identification but it is necessary to address because voice descriptions are often requested before a

lineup identification occurs (Mickes & Wixted, 2015).

Misinformation Effect and Source Confusion

Distortions in memory can occur when influences from outside sources change the witness's initial perception. Loftus (1975) and, more recently, Mori and Kishikawa (2014), found that memory for visual stimuli was impacted after a discussion with a co-witness presented conflicting information. Although the initial event was witnessed visually, the subsequent verbal discussion about the event changed the witness's memory of the event. The misinformation effect is commonly investigated in eyewitness research but less so in earwitness research although identification inaccuracy can result in both instances.

Post-event misinformation presented after hearing an unfamiliar male and female engage in conversation reduced correct identification in a target-present lineup (Smith & Baguley, 2014). After hearing the conversation, witnesses read information stating that either the male or female had a high-pitched voice or a neutral voice. After being given the misleading information of a higher-pitched voice, witnesses rated both voices as having higher pitch levels. Identification accuracy in the target-present voice lineup was above chance (37.5%) but still very low; however, providing verbal recall of the conversation slightly improved identification accuracy.

Memory is susceptible to misinformation from source confusion and may lead to conflicting information when identification is delayed due to a long retention interval (Zaragoza, Belli, & Payment, 2013). At times, it can be several years later when the witness must try to recognize a voice they initially heard. When sources of information are similar, it is very difficult to distinguish between the correct and incorrect source (Mitchell, Johnson, & Mather, 2003).

There is a strong possibility that they may have forgotten everything about the speaker and may have even forgotten about the event. Yet, in the moment, the witness may feel compelled to select an individual from a lineup and possibly testify in court. The passage of time and possible pressure to make an identification can result in their making an inaccurate identification. With the slightest chance that the wrong person can be selected and subsequently incarcerated, more legal procedures need to be implemented to reduce erroneous identification. This thesis highlights the impact of memory interference on material content and examines the witness's ability to discriminate between original and altered content.

2.3 Forensic Impact

Voice Identification Parades

In events where witnesses or victims heard a perpetrator but did not obtain visual verification, voice parades (UK) or lineups (U.S.) using voice samples are conducted to help witnesses or victims attempt to determine the perpetrator's identity (Hollien, 2012). Similar to eyewitness lineups, voice parades conducted in England and Wales present witnesses with nine voice samples including the perpetrator's voice alongside eight comparable voice samples (foils). Each voice sample must be one minute in length and the witness must listen to each voice at least once before making an identification (Home Office, 2003). The guidelines also suggest that the voice parade be performed within 4-6 weeks after the event to reduce memory interference or decay.

In the United States, there is more variation on how voice lineups are conducted. Hollien (2012) recommended a set of standards to adopt. He suggested that six to eight voices including the perpetrator and foils, should be presented in sets of 20-25 voice samples. Each voice sample should be one to two minutes long and the

witness must listen to all the voices in each trial set of voices before selecting a suspect from the trial or choosing not to make a selection. There are some similarities between Hollien's suggested guidelines and the UK guidelines, however, the development of standardized lineup procedures in the U.S. is still ongoing.

A set of guidelines recommended by Wells et al. (2020) offers more guided suggestions for lineup procedures. The guidelines were produced for visual lineups but many recommendations could apply to speaker identification with some modifications. The recommendations are:

1. **Prelineup Interview Recommendation** – to get the witness's description of the suspect and more details surrounding the crime.
2. **Evidence-Based Suspicion Recommendation** – law enforcement should have strong evidence to suspect that a suspect is guilty before including them in the lineup
3. **Double-Blind (or Equivalent) Recommendation** – neither the person conducting the lineup nor the witness should know who the suspect is
4. **Lineup Fillers Recommendation** – the lineup should only have one suspect and at least five foils similar to the suspect
5. **Prelineup Instructions Recommendation** – the instructions should not give any clues about the suspect. The instructions should indicate: (a) the administrator is blind to the lineup, (b) the suspect may or may not be present, (c) witnesses can say that they “don't know”, (d) witness will say how confident they are in their selection, and (e) continue the investigation if the witness did not make an ID.
6. **Immediate Confidence Statement Recommendation** – the confidence rating should be given immediately after the witness's selection

7. **Video-Recording Recommendation** – the whole procedure should be recorded
8. **Avoid Repeated Identifications Recommendation** – do not present the same suspect to the same witness
9. **Showups Recommendation** – avoid showups and try to conduct a lineup

In laboratory experiments, the lineup may be either target-present, meaning that the perpetrator's voice is among those in the lineup or, target-absent, where the perpetrator's voice is absent from the lineup. By employing methods analogous to those used in eyewitness lineups, voice lineups can result in an identification being made. In a target-present voice lineup, the witness can correctly choose the perpetrator (correct identification, or hit), select a filler voice, or foil (false alarm), or reject the correct voice (miss). In a target-absent lineup, the witness can reject the foil voice (correct rejection) or select the foil voice (false alarm) (Kneller, Memon, & Stevenage, 2001).

Current voice lineup procedures contain weaknesses that undermine the reliability of earwitness identification. Research surrounding the lineup presentation suggests that sequential presentations are preferable to simultaneous arrays in visual identifications because they reduce the likelihood of misidentifications in target-absent lineups (Stebly, 1997). The application of sequential and simultaneous lineups is not easily transferable to speaker identification but this thesis will offer an explanation of voice lineups and how they are effective in earwitness identification.

2.4 Literature Summary

This chapter focused on speaker identification and the various factors that affect a witness's performance. Earwitness identification goes beyond merely hearing a voice and making an identification. Previous research has investigated the impact of

various factors like age, gender, and voice familiarity that can reduce earwitness identification accuracy. Acoustical irregularities can also influence the distinctiveness of a voice when making an auditory identification. This chapter explored the memory system that most affects the storage of sounds and word content, examined the memory processes involved, how encoding issues may occur, and explored challenges with retrieval that can lead to misidentification during voice lineups. This chapter further explored how voice lineups are conducted and discussed lineup policies in the UK and recommended guidelines in the U.S. A brief summary in Table 2-1 will further encapsulate the factors that affect identification performance based on current empirical evidence. The present thesis will examine and discuss the following research questions:

- (1) How well do witnesses remember voices?
- (2) Does speaker identification accuracy vary with the gender of the speaker?
- (3) Does speaker identification accuracy vary when the witness is presented with a new voice or new phrase?
- (4) Does speaker familiarity or confidence ratings predict speaker identification accuracy?
- (5) How well do witnesses recall details of a crime?

This thesis will examine the extensive research in the aforementioned areas of earwitness identification and evaluate recent technological advances that offer a new perspective on resolving future misidentification issues and provide more substantial evidentiary value.

Table 2-1*Overview of factors that impact speaker identification performance*

Factor	Performance (Voice Identification)
Familiarity	Familiar voices - ↑ accuracy performance Unfamiliar voices - ↓ accuracy performance
Accent	Familiar - ↑ accuracy performance Unfamiliar - ↓ accuracy performance
Language	Familiar - ↑ accuracy performance Unfamiliar - ↓ accuracy performance
Gender	Male listeners – mixed results for ID of male/female speakers Female listeners - ↑ for male/female speakers but mixed results overall for ID
Age	Children - ↓ for very young group but mixed results for older children Adults - ↑ accuracy performance for ages 20-40
Exposure Time	Short duration - ↓ accuracy performance but mixed results Long duration - ↑ accuracy performance
Retention Interval	Short duration - ↑ accuracy performance Long duration - ↓ accuracy performance but mixed results
Confidence	Mixed results on relationship between accuracy and confidence

Chapter 3 – Experiment 1, Pilot Studies 1 and 2, Experiment 2

3.1 Introduction

How well do listeners remember voices?

An earwitness's ability to remember a voice can impact admissible evidence for legal prosecution. Previous research has shown that listeners are able to recognize familiar voices but the uncertainty remains as to how well they can recognize voices of unfamiliar speakers that they have heard once (Yarmey et al., 2001). Several factors are involved in earwitness identification to determine how well an earwitness can identify a speaker's voice and whether their testimony is admissible evidence for legal prosecution. Witnesses who are familiar with a speaker's voice may still struggle to make a correct identification. Ladefoged and Ladefoged (1980) tested Ladefoged's own ability to identify familiar voices and was able to correctly identify 31% of voices speaking the word, "Hello," but failed to recognize the voice of his mother in the process. Given that voice samples of 2 seconds have been correctly identified at a rate above chance (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012), arguably a single word is sufficient to offer specific characteristic markers for the witness to encode the voice.

The length of a word or a series of words is equivalent to the length of exposure to a speaking voice. The length of exposure to a voice may determine whether the information is encoded. Short, two-second samples successfully identified voices above chance; however, research has shown that a longer exposure duration is more likely to increase identification accuracy (Yarmey, 2012). Kerstholt et al. (2004) analyzed the effect of exposure time on accuracy and did not find that participants performed much better with a longer exposure duration of 70s (46%) than a shorter exposure duration of 30s (38%). Overall, participants who viewed a voice lineup one

week after exposure to the target voice, performed better when they were exposed to the voice for a longer duration than a short duration. Kerstholt et al. (2004) explained that the overall performance was positive and it was likely that participants would only identify an innocent person 9% of the time.

Multiple exposures to a voice can impact identification accuracy. Deffenbacher et al. (1989) exposed participants to a voice heard one time or multiple times over a period of three days. The results showed that, although participants were able to identify the voice, identification accuracy was low. Yarmey and Matthys (1992) compared one time voice exposures to repeated voice exposures. Participants heard a single-voice speech for 18 seconds, 36 seconds, 120 seconds, or 6 minutes. Participants were exposed to the voice either at one time, for two exposures (half of sample length per exposure), or three exposures (1/3 of sample length per exposure). After hearing the voice, the participants were either given an immediate lineup, a lineup after 24 hours, or after one week. They were told that the suspect may or may not be present in the lineup. Six voices were presented in the target-present lineup as well as the target-absent lineup. In the target-present condition, performance was most robust when participants were given a voice sample of 120 seconds but it also produced the highest number of false alarms. Hit rates were higher when the voice was presented twice but there was no difference between identification accuracy of a single voice exposure to the three-time voice exposure. In the target-absent condition, longer durations increased false alarms. The 6-minute exposure still resulted in false alarms in the target-absent condition. The overall results suggest that accurate identification is challenging and increased voice exposure does not necessarily improve performance (Yarmey & Matthys, 1992).

In addition to the length of exposure and voice variability, individuals must

contend with the retention interval length. The retention interval is the length of time between witnessing the criminal event and making an identification in a voice lineup (Sherrin, 2015). Research has shown that accuracy rates vary based on the length between the initial voice sample exposure and the recognition period (Sherrin, 2015). Retention intervals in laboratory experiments can be immediate, or typically, a duration of hours, days, or weeks.

Kerstholt et al. (2004) examined retention intervals of short and long durations. Participants heard eight voice samples in target-present and target-absent lineup conditions. They participated in a voice lineup either immediately or asked to return a week later. Next, participants heard six voices in the target-present and target-absent lineups. After they heard all six voices, they determined whether the target voice was presented. If they were unsure, they were forced to choose whether the target was present or absent in the lineup. After providing their answer, they indicated on a seven-point Likert scale how confident they were in their answer. Performance in the target-absent condition was low as participants identified a foil voice incorrectly as the target (51%); however, in the target-present condition, participants accurately identified the correct voice in the lineup (42%) rather than selecting a foil voice (24%). Overall, participants who viewed the lineup one week after the learning phase performed better (47%) than the participants who immediately viewed the lineup (38%). These results are in contrast to most studies that suggest a longer retention interval decreases accuracy. As previously mentioned, Charles Lindbergh identified Bruno Hauptmann's voice three years after the initial exposure to the perpetrator's voice. In real life experiences, retention intervals can extend as long as months or several years.

The gender of the witness and the speaker can also impact identification

accuracy. Previous research has found a gender-bias where female witnesses identify female speakers better than males and vice versa (Roebuck 1993). When exposed to familiar voices, female listeners accurately identified male and female voices better than male listeners (Skuk & Schweinberger, 2013). Conversely, Cook & Wilding (1997b) tested male and female participants in voice identification accuracy. Participants listened to an audio tape of one male and one female voice. They were asked to return a week later to identify the target speaker out of a six-voice lineup. There were no significant differences among the male and female participants when identifying male and female voices in the lineup. Yarmey and Matthys (1992) found that in the longest exposure time of six minutes, female participants performed worse than males. However, there were no significant gender differences between male and female participants and their accuracy scores for male and female voices.

In real-life situations, it is impossible to determine if the lineup is a target-present or target-absent lineup. In the laboratory, these variables are much easier to control (Orchard & Yarmey, 1995). In a target-present lineup, a witness may correctly identify the target voice (hit), incorrectly identify a voice (false alarm), or determine the target voice is not presented (miss). In a target-absent lineup, a witness may reject an incorrect voice (correct rejection) or select an incorrect voice (false alarm) (Kneller et al., 2001).

There is some debate as to the most optimal choice to present an eyewitness lineup. In eyewitness identification, the witness may be presented with a simultaneous lineup or sequential lineup. In a simultaneous lineup, the witness is presented with several faces at the same time (Wells et al., 1998). In a sequential lineup, the witness is presented one face at a time. In the U.S., the witnesses are asked to determine if the presented face is the perpetrator before moving onto the next face. This type of

sequential lineup reduces the possibility of relative judgment where the witness will compare each presented face to the other faces in the lineup. In the UK, witnesses are presented with the faces at least twice before making an identification (Brewer & Palmer, 2010). In target-present lineups, witnesses identified the perpetrator at a higher rate when they were able to view all the faces more than once (65%) compared to witnesses who had to determine whether the face was or was not the perpetrator after each face was presented (36%) (Valentine, Pickering, & Darling, 2003).

The lineup procedures used for eyewitness identification may not achieve the same accuracy rates in earwitness identification. Valentine et al. (2003) showed that making an identification after seeing all the presented faces in a simultaneous lineup led to more correct identifications. While some eyewitness procedures may be applicable to earwitness identification, voice lineups are either serial or sequential presentations (Smith et al., 2020). A sequential lineup presents voices in sequential order (i.e. one voice followed by another until the end of the lineup). Similar to the U.S. sequential eyewitness lineup, the witness must decide to select that voice or move on to the next voice. In contrast, a serial lineup requires the witness to make an identification at the end of the lineup after hearing all the voice samples. Smith et al. (2020) found that participants accurately identified voices in sequential lineups ($M = 39.13$, $SD = 49.90$) better than serial lineups ($M = 16.67$, $SD = 38.07$) in target-present conditions as well as in target-absent conditions ($M = 17.39$, $SD = 38.76$ and $M = 9.52$, $SD = 30.08$, respectively). The results suggest that different strategies are needed for visual and auditory modalities (Yarmey, 1995).

Voices can vary in tone, pitch, emotion, and listening environment and make identification difficult. Acoustical variability like pitch, tone, and speaking rate can impact speaker identification. Previous research has found that variability in voices,

rather than the duration of target voice samples presented, can be a determining factor in speaker identification inaccuracy (Sherrin, 2015). A witness creates markers during the initial encoding process that marks aural characteristics like pitch, tone, and speaking rate (Mullennix et al., 2010). When those markers are changed during the identification process, it is difficult for the witness to match the voice lineup sample to the initial encoded voice (Dewhurst, Bould, Knott, & Thorley, 2009).

Disguised voices or whispered voices heard during the initial exposure have proven difficult to recognize when presented with a “normal” voice during the identification period (Orchard and Yarmey, 1995, as cited in Kerstholt et al., 2004). Perpetrators can use tonal changes like emotionality and whispering to disguise their voices. When a witness is initially presented with a particular tone of voice, the voice is retained in the long-term memory. During the identification process, the tone of voice should match the same tone that was initially encoded. Accents and language can also impact speaker identification when they differ from the witness’ accent or language (Stevenage et al., 2012). Witnesses are more likely to correctly identify a speaker's voice when the accent or language is familiar rather than unfamiliar.

Age plays an important factor in earwitness identification because young to middle-aged adults tend to outperform children and elderly witnesses in speaker identification. However, children as young as five can still correctly identify a familiar voice (Yarmey, 2012). In the present thesis, participants’ ages were recorded for descriptive purposes but this thesis does not further address age group differences or the impact of age on identification accuracy.

How well do listeners remember what was said?

Palmeri et al. (1993) analyzed listener performance on word identification. In Experiment 1, they presented listeners with 140 monosyllabic test words that were

repeated by 1, 2, 6, 12, or 20 voices (equal number of male and female voices per presentation) and the listeners had to discriminate between old and new words. In the multiple-speaker test groups, listeners performed better with words they previously heard when spoken in the same voice rather than a different voice. Accuracy rates for the single-speaker group showed similar results as the same-voice presentation in the multiple speaker group. Accuracy was higher in the same speaker-same gender group than the different speaker-same gender group and the different speaker-different gender group, but their recognition performance was not affected by the increase of speaking voices. However, the increase in the lag times between word presentations reduced accuracy rates. The results explain that voice characteristics serve as retrieval cues for spoken voice codes that are retained in long-term memory.

In Experiment 2, the methodology was the same, but the listeners were presented with 84 monosyllabic words and had to answer if the word was new or old. For the old responses, the listeners had to determine if the word was presented in the original, same voice or a different voice. Similar to Experiment 1, they found that the increase in lag time decreased accuracy performance. Overall, listeners performed best in the same speaking voice group regardless of the speaker's gender; however, listeners were more accurate with words presented in a different voice by a different gender than voices presented in a different voice by the same gender.

Earwitness identification is flawed and a number of factors exist that impact how well a witness can identify a speaker's voice. We attempted to replicate the experiment by Palmeri et al. (1993). The aim of Experiment 1 was to examine how well listeners identify words presented in the same voices they previously heard and how well they identify voices speaking the same words that they heard previously. Within this context, we examined whether accuracy changed based on the speaker's

gender or when the listener was presented with a new word or a new voice.

Similar to Palmeri et al. (1993), we used monosyllabic words for both conditions and we analyzed whether memory accuracy varied with the gender of the speaker in each condition. Contrary to voice lineup recommendations, the monosyllabic voice samples were short in duration rather than the suggested length of one-minute long samples. Palmeri et al. (1993) found that identification performance accuracy was not affected by the number of samples, the length of the samples, or the gender the speaker. Our attempt to replicate those findings would support a review of current voice parade procedures and potentially effectuate new policy measures.

Pilot studies 1 and 2 were conducted to determine whether participants were attending to the stimuli or making arbitrary selections above chance. In Experiment 2, we included a written crime scenario involving a Robber and a Shop Attendant. The rationale was to analyze how well listeners remembered voices they heard in a violent scenario context. In scenarios where physical violence has occurred, witnesses have recounted more accurate details of the event than when violence did not occur (Pajón & Walsh, 2017).

3.2 Experiment 1

The purpose of this experiment was to assess a baseline for memory recognition for words and voices in the most basic form before analyzing the complexities of attention, exposure time, and retention duration. The aim of the experiment was to analyze how well listeners remembered voices and words and whether accuracy varied based on the speaker's gender. We focused on memory recognition of auditory voice stimuli no longer than two seconds in duration because previous research has shown that witnesses can still recognize voices after hearing them for a short duration (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012).

Participants were randomly allocated to one of two conditions and asked to distinguish between auditory words or voices they previously heard and newly introduced words and voices. After listening to the stimuli during the learning phase, participants completed a brief filler task that lasted no longer than 10 seconds before being presented with the recognition test. We predicted that auditory words would be recognized at a higher rate of accuracy than voices (H1) because semantic memory is focused on the meaning of the words which promotes a deeper level of processing than phonemic sounds associated with the spoken words (Holt, 2019). We also anticipated that both words and voices would be recognized with greater accuracy when spoken by female voices than male voices (H2).

3.2.1 Method

Design

The experiment was a repeated measures design. Each condition was analyzed separately as the manipulations for words and voices were not interchangeable. In the auditory word recognition condition, listeners heard forty words spoken in various voices and were asked if the word they heard was previously presented or a new word. The response of either “Old” or “New” word was the dependent variable and the independent variable tested was the gender of the speaker (male or female). In the speaker identification condition, listeners heard forty voices and were asked if the speaking voice they heard was previously presented or a new voice. The response of either “Old” or “New” voice was the dependent variable and the independent variable tested was the gender of the speaker (male or female).

Participants

Fifty-six adults and undergraduate students (36 females and 20 males aged between 18-53, $M = 24.25$, $SD = 7.888$) at the City, University of London, participated

in the study in exchange for departmental credit or monetary compensation for transportation costs. All first-year City, University of London Psychology undergraduates received departmental credit. Second- and third-year undergraduates and adults recruited through City, University of London SONA Online participant management database received £5. All participants were fluent English speakers, and none reported any hearing impairments that would have prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

Speech recordings were made in the anechoic chamber of the Department of Phonetics and Linguistics, University College London (“UCL”) using a Brüel & Kjær sound level meter. The glottal activity was measured using an electro-laryngograph, and recordings were made to Digital Audio Tape (“DAT”) at a sampling rate of 44.1 kHz (Markham & Hazan, 2002). Each word was a separate audio file that was uploaded into E-Prime computer software and presented on a PC computer terminal. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 decibels (“dB”).

Procedure

Of the 56 participants, 29 participants were randomly assigned to the auditory word recognition condition and 27 were assigned to the speaker identification condition. Each experiment was divided into three parts, involving a learning session, a ten second visual filled task, and a word recognition or speaker identification test session. Participants performed each test on a desktop computer in a research cubicle at City, University of London.

Before starting the learning session, participants read an information sheet that

explained their rights to participate and signed a consent form. In the first session, or learning session, participants listened to forty monosyllabic words presented in succession. Five male speakers and five female speakers spoke four words each, and the words were presented in random order for each participant. At the conclusion, of the learning session, the experimenter presented the participant with a visual filled task of viewing a paper picture illusion for ten seconds before proceeding to the recognition test.

Participants in the auditory word recognition condition were instructed that they would listen to forty monosyllabic words and after each word, they were required to answer if they heard the word in the previous learning session or indicate if the word was new. During the recognition testing session, forty monosyllabic words were presented individually. Twenty of the words were initially presented in the learning session and twenty were new words. All words were presented in the same speaking voices used in the learning session. Participants were required to select the letter “A” for an old word (previously presented during the learning the session) and the letter “L” for a new word (newly presented in during the testing session).

Participants in the speaker identification condition were instructed that they would listen to forty monosyllabic words and after each word, they were required to answer if they heard the speaker’s voice in the previous learning session or indicate if it was a new voice. During the testing session, forty monosyllabic words were presented in random order. The words presented in the testing session were the same words used in the learning session. The voice samples consisted of twenty voices initially presented in the learning session and twenty new voices. Participants selected the letter “A” for an old voice (previously heard in the learning session) and the letter “L” for a new voice (newly presented during the testing session).

Participants in both conditions were not given prior warning of the recognition testing session before the start of the testing session. After completion of the testing session, the participant was thanked for his or her participation and debriefed on the aim of the experiment. All scores were tallied based on the signal detection measures.

For both conditions, the participants' responses were tallied as 0 or 1 based on signal detection measures (Stanislaw & Todorov, 1999). Participants received a score of 1 for a "hit." A hit was defined as correctly recognizing the "target" voice which was the voice they previously heard in the learning session. They also received a score of 1 for correctly rejecting the "non-target" voice which was the new voice introduced in the testing session. Comparatively, they received a score of 0 for missing the target voice or incorrectly identifying a non-target voice as an "old" voice. The hit rates and correction scores were calculated based on the total response scores divided by a total number of voice samples. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1. Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding "old" and a positive c score indicated the likelihood of responding "new."

3.2.2 Results

Response Scores – Auditory Word Recognition

In the auditory word recognition condition, we found that hit rates were slightly higher for words spoken in male voices (69%) than words spoken in female voices (63%) but false alarms were consistently lower for both speaker genders (31% for male speakers and 32% for female speakers).

The response bias, c , for words spoken in male voices ranged from -0.994 to 0.767 ($M = -0.033$, $SD = 0.472$) and for words spoken in female voices ranged from -1.247 to 1.129 ($M = 0.085$, $SD = 0.528$). Participants were more likely to respond “old” to words spoken in male voices and “new” to words spoken in female voices.

Table 3-1

Mean c response bias and d' scores for Auditory Word Recognition

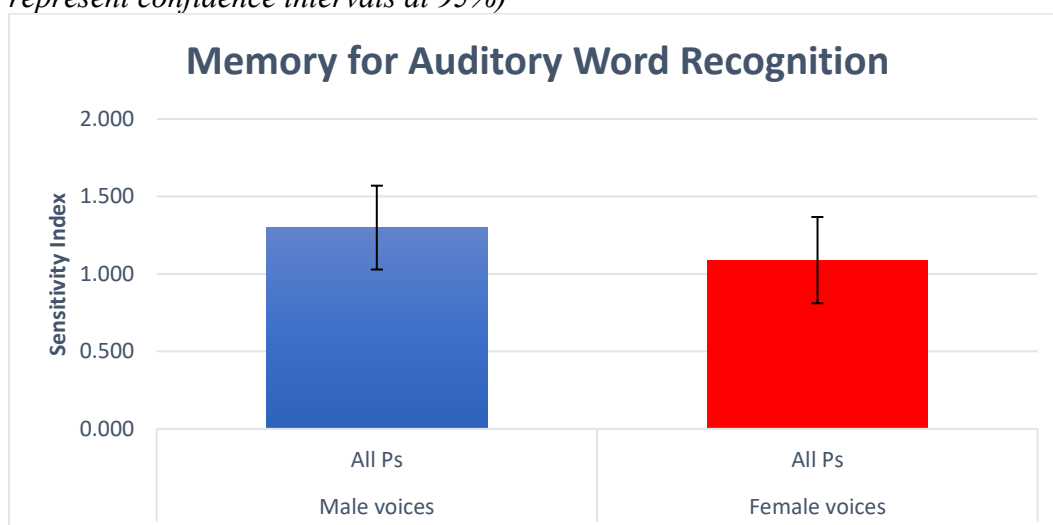
	Male voices	Female voices
Mean	1.299	1.089
SD	0.743	0.764
SE	0.138	0.142
c response bias	-0.033	0.085

Identification Accuracy Scores

A one-way ANOVA was conducted with speaker gender (male d' prime score, female d' prime score) as a within-subjects factor. The results did not show a significant effect of speaker gender $F(1,28) = 1.483$, $p = .234$, $\eta_p^2 = .050$.

Figure 3-1

Mean d' scores for auditory word recognition for speaker gender (error bars represent confidence intervals at 95%)



Response Scores – Speaker Identification

In the speaker identification condition, we found that hit rates were higher for male voices (72%) than female voices (64%). However, false alarms were higher for male voices (61%) than female voices (37%).

The response bias, c , for male voices ranged from -1.769 to 1.391 ($M = -0.501$, $SD = 0.615$) and for female voices ranged from -0.767 to 1.062 ($M = 0.025$, $SD = 0.434$) which showed that participants were more likely to respond “old” to male voices and “new” to female voices.

Table 3-2

Mean c response bias and d' scores for Speaker Identification

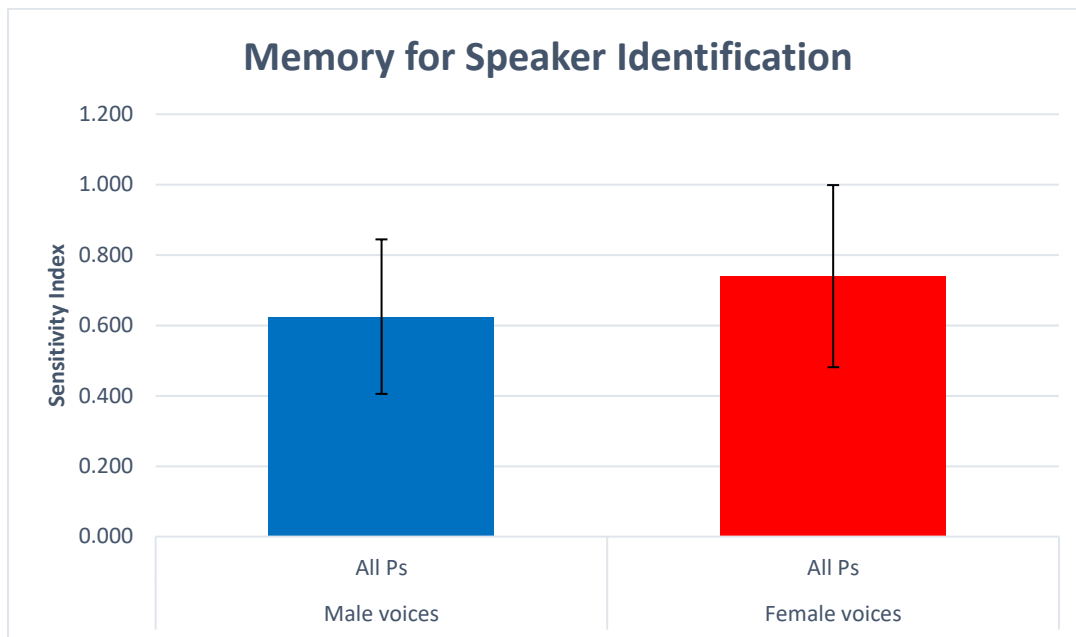
	Male voices	Female voices
Mean	0.625	0.740
SD	0.584	0.687
SE	0.112	0.132
c response bias	-0.501	-0.025

Identification Accuracy Scores

A one-way ANOVA was conducted with speaker gender (male d' prime score, female d' prime score) as the within-subjects factor. The results did not show a significant effect of speaker gender $F(1,26) = 0.401$, $p = .532$, $\eta_p^2 = .015$.

Figure 3-2

Mean d' scores for speaker identification based on speaker gender (error bars represent confidence intervals at 95%)



3.3 Pilot Study 1

The pilot study was conducted to determine whether participants were attending to the stimuli or making arbitrary selections at a rate above chance. We examined the recognition accuracy of short duration voice samples comparing male and female voices. The same voice samples from Experiment 1 were presented; however, the presentation of the voice samples was changed to a repetitive format. To simulate an increase in voice exposure, four voice samples for each speaker were presented in a sequence of one after the other rather than one sample per speaker presented in random order like Experiment 1. We predicted that repetitive exposure to a voice would increase the recognition accuracy rates during the speaker identification test (H1), and listeners would identify female voices more accurately than male voices (H2).

3.3.1 Method

Design

This was a short pilot study to determine if participants were attending to the voice samples. The experiment was a repeated measures design. Like Experiment 1, listeners heard forty voices and the response of either “Old” or “New” voice was the dependent variable and the independent variable tested was the gender of the speaker (male or female).

Participants

Eleven undergraduate students at the City, University of London and non-City affiliated adults (9 females and 2 males aged between 19-53, $M = 30.09$, $SD = 8.803$) participated in the study in exchange for departmental credit or monetary compensation for transportation costs. All first-year City, University of London Psychology undergraduates received departmental credit. Second- and third-year undergraduates, and adults recruited through City, University of London SONA online participant management database received £4. All participants were fluent English speakers, and none reported any hearing impairments that would have prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

The speech recordings were the same recordings used in Experiment 1. Each word was a separate audio file that was uploaded into E-Prime computer software and presented on a PC computer terminal. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 dB.

Procedure

The twenty-minute pilot study was divided into three parts, involving a learning session, a ten second visual filler task, and a speaker identification test

session. In the learning session, forty monosyllabic words were presented by five male speakers and five female speakers, each speaking four different words presented in consecutive blocks for each speaker. The same voice stimuli from Experiment 1 were presented.

Before beginning the study, participants read an information sheet that explained their rights as a participant and signed a consent form. Prior to beginning the learning session, participants were instructed that they would listen to forty words. In the learning session, participants listened to forty monosyllabic words played in succession. Five male speakers and five female speakers spoke four words each. Presentation of the blocks was counterbalanced; half of the participants heard a list that began with a female speaker, while the other half heard a list that began with a male speaker.

Before the speaker identification test, participants were instructed that they would listen to forty monosyllabic words spoken in various voices. After each word, they were required to answer whether they heard the speaker's voice in the previous learning session by selecting the letter "A" for old voice or the letter "L" if the voice was a new voice. They were not given prior notice of the recognition testing session before the start of the testing session. During the session, forty monosyllabic words were presented individually. The voices consisted of twenty voices originally presented in the learning session and twenty new voices. After completion of the testing session, the participant was thanked for his or her participation and debriefed about the aim of the pilot study.

The hit rate and correction scores were calculated based on the total response scores divided by a total number of voice samples. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1.

Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding “old” and a positive score indicated the likelihood of responding “new.”

3.3.2 Results

Response Scores

We found that participants performed better in memory recognition for male voices (72%) than for female voices (69%). However, there were higher false alarm rates for male voices (62%) than female voices (29%).

The response bias, c , for male voices ranged from -2.495 to 2.510 ($M = -0.601$, $SD = 0.410$) and for female voices ranged from -0.903 to 0.547 ($M = 0.012$, $SD = 0.431$) which showed that participants were more likely to respond “old” to male voices and “new” to female voices.

Table 3-3

Mean d' scores for Voice Identification in Repeated Exposure

	Male voices	Female voices
	All Ps	All Ps
Mean	0.230	1.153
SD	1.474	0.892
SE	0.445	0.269
c response bias	-0.601	0.012

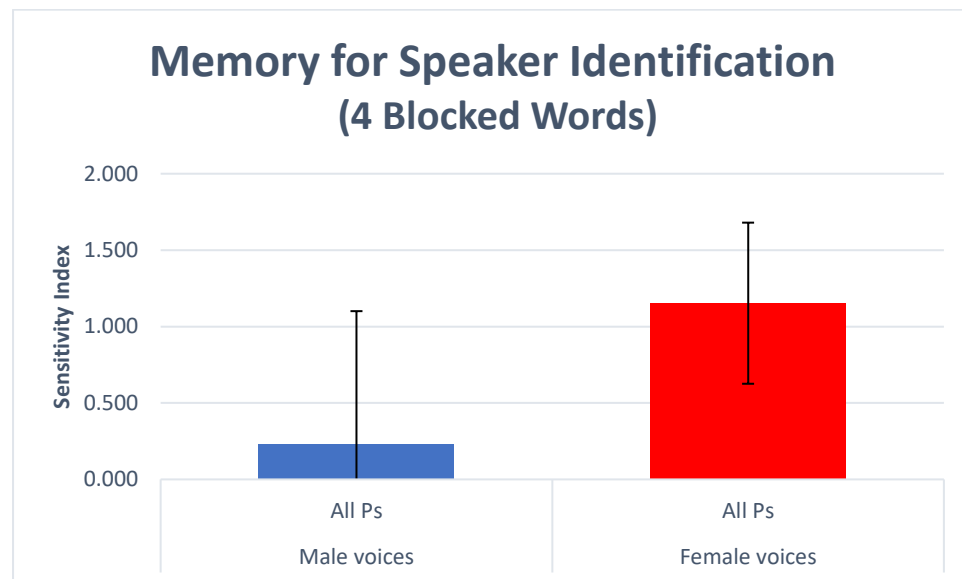
Identification Accuracy Scores

A one-way ANOVA with speaker gender (male d' prime score, female d' prime score) as a within-subjects factor did not show an effect of speaker gender $F(1,10) =$

2.660, $p = .134$, $\eta_p^2 = .210$. Taking into account that this was a pilot study, the small sample size is not an adequate reflection of gender differences in speaker identification.

Figure 3-3

Mean d' scores for speaker identification based on speaker gender (error bars represent confidence intervals at 95%)



3.4 Pilot Study 2

Similar to Pilot Study 1, the pilot study was conducted to determine whether participants were attending to the stimuli or randomly selecting voices at a rate above chance. The same voice samples from Pilot study 1 were presented here, however, the presentation of the voice samples was changed to a two-word or four-word sequential voice sample format to reduce order effects. Exposure was the same as presented in Pilot Study 1 but the voice samples for each speaker were presented in a repetitive sequence of either two or four samples rather than one sample per speaker as presented in Experiment 1. We predicted that overall false alarm rates would be reduced and hit rates would increase in all samples (H1) and listeners would identify female voices more accurately than male voices (H2).

3.4.1 Method

Design

The experiment was a repeated measures design. Like Experiment 1, listeners heard forty voices and the response of either “Old” or “New” voice was the dependent variable and the independent variable tested was the gender of the speaker (male or female).

Participants

Eighteen undergraduate students (17 females and one male aged between 18-22, $M = 18.71$, $SD = 1.105$) at the City, University of London, participated in this twenty-minute study in exchange for departmental credit or monetary compensation for transportation costs. All first-year City, University of London Psychology undergraduates received departmental credit. Second- and third-year undergraduates recruited through City, University of London SONA Online participant management database received £4. All participants were fluent English speakers, and none reported any hearing impairments that would have prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

The speech recordings were the same recordings used in Experiment 1. Each word was a separate audio file that was uploaded into E-Prime computer software and presented on a PC computer terminal. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 dB.

Procedure

The twenty-minute experiment consisted of one session. In the session, forty

monosyllabic words were presented by five male speakers and five female speakers, each speaking four words. The speaker's voices were presented in sequences of four of the same voice or two of the same voice and alternated between groups of four voices followed by groups of two voices (e.g. Voice 1A, 1B, 1C, 1D, Voice 2A, 2B, Voice 3A, 3B, 3C, 3D, etc.). Before beginning the experiment, participants read an information sheet that explained their rights as a participant and signed a consent form. Prior to beginning the session, participants were instructed that they would listen to forty words and instructed to indicate whether the voice they heard was the same or different voice than the voice immediately preceding that voice. Five male speakers and five female speakers spoke four words each, and the words were presented in random order for each participant. After completion of the session, the participant was thanked for his or her participation and debriefed on the aim of the experiment.

The hit rate and correction scores were calculated based on the total response scores divided by a total number of voice samples. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1. Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding "old" and a positive score indicated the likelihood of responding "new."

3.4.2 Results

Response Scores

We found that all participants performed slightly better in speaker identification for male voices (86%) than for female voices (84%). We found, overall,

that there were higher false alarm rates for male voices (32%) than female voices (18%).

The response bias, c , for male voices ranged from -0.852 to 0.066 ($M = -0.377$, $SD = 0.316$) and for female voices ranged from -0.651 to 0.426 ($M = -0.024$, $SD = 0.296$) which showed that participants were more likely to respond “old” to male voices and female voices.

Table 3-4

Mean d' scores for Voice Identification in Repeated Voice Exposure

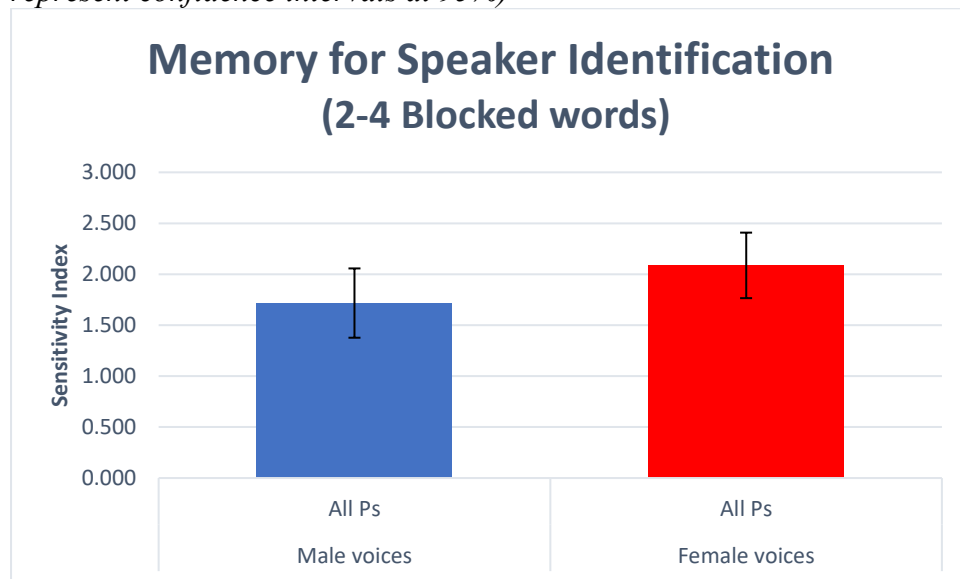
	Male voices	Female voices
Mean	1.717	2.087
SD	0.715	0.677
SE	0.174	0.164
c response bias	-0.377	-0.024

Identification Accuracy Scores

A one-way ANOVA with speaker gender (male d' prime score, female d' prime score) as a within-subjects factor did show a significant effect of speaker gender $F(1,16) = 5.231$, $p = .036$, $\eta_p^2 = .246$. Taking into account that this was a pilot study, the small sample size is not an adequate reflection of gender differences in speaker identification.

Figure 3-4

Mean d' scores for speaker identification based on speaker gender (error bars represent confidence intervals at 95%)



3.5 Experiment 2

Previous research has shown that eyewitnesses correctly recalled more information about an event involving physical violence than a non-violent event (Pajón & Walsh, 2017). Additionally, obscene and explicit material is typically remembered with more accuracy than neutral material (Leander, Granhag, & Christianson, 2005). Therefore, it is likely that a crime scenario involving a distressing robbery may lead to more attention to detail and careful encoding of information for detailed retrieval later. The aim of Experiment 2 was to present a violent robbery scenario and examine how well listeners identify the robber speaking the same words that they heard previously. Within this context, we examined whether accuracy changed based on the robber’s gender or when the listener was presented with a new voice. We predicted that speaker identification rates would be more accurate for the “Male Robber” in the crime scenario (H1). We also predicted that listeners would identify female voices better than male voices (H2).

3.5.1 Method

Design

The experiment was a between-subjects design. Participants read one of two on-screen crime scenarios and took part in either the “Female Robber” condition or the “Male Robber” condition. Listeners heard forty voices and the response of “Robber” or “Attendant” based on the scenario or a new voice was the dependent variable and the independent variable tested was the gender of the speaker (male or female).

Participants

Nineteen undergraduate students at the City, University of London and non-City affiliated adults (18 females and one male aged between 18-25, $M = 18.79$, $SD = 1.652$) participated in the study in exchange for departmental credit or monetary compensation. All first-year City, University of London Psychology undergraduates received departmental credit. Second- and third-year undergraduates, and adults recruited through City, University of London SONA Online participant management database received £4. All participants were fluent English speakers, and none reported any hearing impairments that would have prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

The speech recordings were same the voice samples presented in Experiment 1. Each word was a separate audio file that was uploaded into E-Prime computer software and presented on a PC computer terminal. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 dB.

The scenario was a false story of a witness observing a robbery taking place at a shop (see Appendix A). The scenario was displayed electronically on a computer PC in the City, University of London laboratory offices.

Procedure

Of the 19 participants, 8 participants (all females) were randomly assigned to the female robber condition and 11 were assigned to the male robber condition (10 females and one male). The twenty-minute experiment was divided into four parts: reading a written scenario, a learning session, a ten-second visual filled task, and a speaker identification test session. Participants were presented with a scenario of a fictitious robbery that occurred in a shop. In the female robber condition, the robber was female and the shop attendant was male. In the male robber condition, the shop attendant was female and the robber was male. In the learning session, forty monosyllabic words were presented by five male speakers and five female speakers each speaking four words presented in random order. The same stimuli voice samples from Experiment 1 were presented.

Before beginning the experiment, participants read an information sheet that explained their rights as an experiment participant and signed a consent form. Prior to beginning the learning session, participants were instructed that they would read a short scenario followed by the audio presentation of forty words. In the learning session, participants listened to forty monosyllabic words played in succession. During the presentation of each word, either “robber” or “attendant” was displayed on the screen based on the respective genders of the attendant and robber in the aforementioned scenario. Five male speakers and five female speakers spoke four words each and the list of words were presented in random order for each participant.

Before the speaker identification test, participants were instructed that they

would again hear forty monosyllabic words spoken in various voices and that for each word, they should indicate whether the voice they heard was that of the “robber” or the “attendant” from the previous learning session or a new voice. Participants made a choice by selecting the letter “A” for “robber” or “attendant” or the letter “L” for a new voice. They were not given prior notice of the speaker identification test session before it began. The voices consisted of twenty voices originally presented in the learning session and twenty new voices. After completion of the speaker identification test, the participant was thanked for his or her participation and debriefed about the aims of the experiment.

The hit rate and correction scores were calculated based on the total response scores divided by a total number of voice samples. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1. Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding “old” and a positive score indicated the likelihood of responding “new.”

3.5.2 Results

Response scores

Female Robber Scenario

We found that participants performed better in memory recognition for male voices (71%) than female voices (56%). Overall, the false alarm rates were higher for male voices (70%) than female voices (26%).

The response bias, c , for male voices ranged from -2.257 to 0.127 ($M = -0.694$, $SD = 0.729$) and for female voices ranged from 0.000 to 0.641 ($M = 0.255$,

$SD = 0.249$) which showed that participants were more likely to respond “old” to male voices and “new” to female voices.

Response scores

Male Robber Scenario

Participants performed better in memory recognition for male voices (62%) than for female voices (55%). We found that participants had higher false alarm rates for male voices (64%) than female voices (60%).

The response bias, c , for male voices ranged from -1.769 to -0.127 ($M = -0.491$, $SD = 0.517$) and for female voices ranged from -0.547 to 1.391 ($M = 0.304$, $SD = 0.574$) which showed that participants were more likely to respond “old” to male voices and “new” to female voices.

Table 3-5

Mean d' scores for Voice Identification in Male and Female Robber conditions

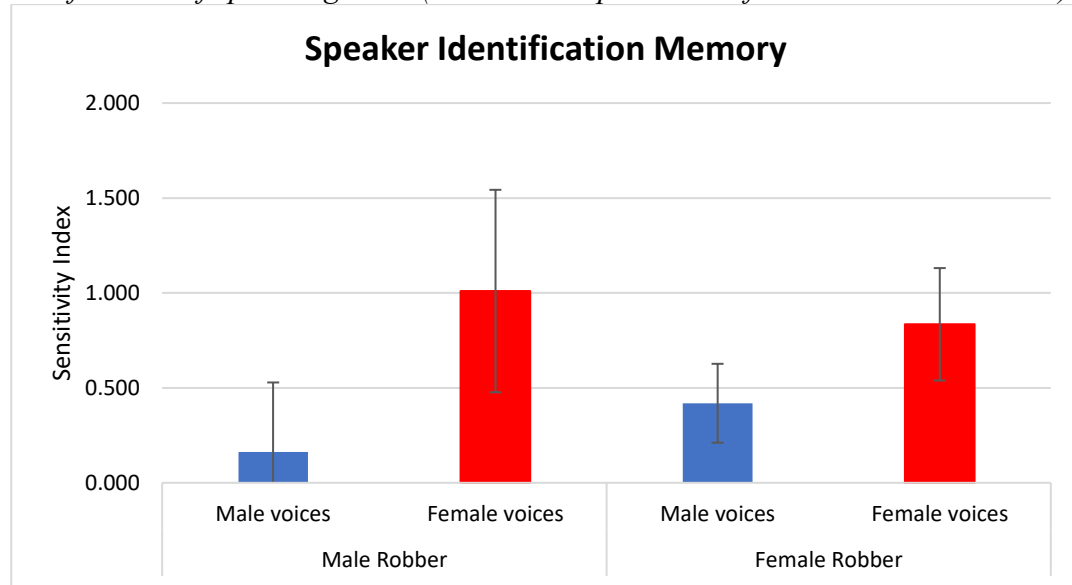
	Male Robber		Female Robber	
	Male voices	Female voices	Male voices	Female voices
Mean	0.162	1.010	0.419	0.835
SD	0.619	0.902	0.299	0.427
SE	0.187	0.272	0.106	0.151
c response bias	-0.491	0.304	-0.694	0.255

A 2X2 mixed model ANOVA with Robber gender as the between-subjects factor (male robber, female robber) and speaker gender (male d' prime score, female d' prime score) as the within-subjects factor showed a significant main effect of speaker gender $F(1,17) = 14.308$, $p = .001$, $\eta_p^2 = .457$. There was no significant effect on Robber gender $F(1,17) = 0.942$, $p = .345$, $\eta_p^2 = .052$ nor a significant interaction

between Robber gender and speaker gender $F(1,17) = 3.249, p = .089, \eta_p^2 = .160$. The small samples are not an adequate reflection of gender differences in speaker identification.

Figure 3-5

Mean d' scores for memory voice identification for Robber gender as a function of speaker gender (error bars represent confidence intervals at 95%)



3.5.3 Discussion

Our results showed a significant effect of speaker gender in both Pilot 2 and Experiment 2. In Pilot 2, participants accurately selected male voices at a higher rate than female voices. In Experiment 2, participants also performed better with male voices than female voices regardless of the gender of the Robber in the scenario. These results do not support the hypotheses that female voices will be more accurately identified than male voices; however, it is critical to note that the small sample sizes for each experiment prevent these results from adequately reflecting any gender differences in speaker identification. The trend of the results indicated that participants accurately recognize old words even when they are spoken in a new voice they did not previously hear but struggled with identifying old voices speaking the

same words they heard in the learning phase. Another trend showed that participants were better at recognizing words spoken in a male voice than a female voice but false alarm rates for male voices were much higher than those for female voices.

The results from Experiment 1 gave us a baseline from which to further our focus in the chapters to follow. Rather than investigating word identification in Pilot studies 1 and 2 and Experiment 2, we focused solely on speaker identification accuracy to determine how well participants recognize male and female voices they have previously heard. The two-second duration voice samples produced high hit rates but also high false alarm rates. It is likely that longer voice samples are likely to produce much higher accuracy rates as previous research has explored (Kerstholt et al., 2004). By starting with a shorter voice sample, we can effectively compare the impact of exposure duration as we progress through the subsequent studies.

Previous research has shown that repetitive exposure to voice samples increases speaker identification accuracy rates (Kerstholt et al., 2004). Although voice samples of a 2-second duration or longer can produce recognition rates above chance, samples longer in duration have been found to increase identification accuracy (Bull & Clifford, 1984, as cited in Yarmey, 1992). In the case of the *State v. Hauptmann* (1935), Charles Lindbergh provided eyewitness testimony stating that he recognized the defendant's voice when he heard him say, "Hey, doc, over here." While it is plausible that short duration voice samples can lead to accurate speaker identification, it can be challenging to prove that shorter samples coupled with an extended retention interval of several years will yield highly accurate identification rates. Unfortunately, the urgency to apprehend a suspect and secure a conviction tends to overlook this flaw. This not only leads to false alarms whereby innocent persons are implicated, but it also promotes ongoing fragility in the law enforcement process and legal system.

Longer duration voice samples have produced an increase in accurate identification, but have also led to higher false alarm rates (Yarmey, 1991). Unfortunately, there is no set sample length to ensure higher recognition accuracy. Researchers have studied voice sample lengths of a few seconds to several minutes to determine what duration may produce improved performance; however, results have varied (Kerstholt, Jansen, Van Amersvoort, & Broeders, 2004). In support of previous research, our voice samples met the minimum duration standard necessary to produce recognition accuracy above chance (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012) but fell short of the recommended length of one minute required in the UK (Home Office, 2003) and 1-2 minutes suggested by Hollien (2012). In the chapters to follow, a further analysis of extended duration time will be discussed.

Simulating real-life conditions whereby short duration crimes occur in a matter of minutes or less, is essential to understanding how accurate earwitnesses are in identifying voices later presented in a voice lineup or parade. Crimes like burglary or assault can occur quickly by perpetrators in disguise with very few words exchanged in the process. The retention interval length has been shown to influence recognition accuracy even when longer duration voice samples have been presented (Clifford, Rathborn, & Bull, 1981). A lengthy voice sample of several minutes may show speaker identification rates at chance level within 24 hours of hearing the samples. Research has shown that over a 1-, 2-, and 3-week retention period, recognition accuracy rates decreased to 9 percent (Clifford, Rathborn, & Bull, 1981). Such results make it difficult to substantiate the accuracy of the Lindbergh testimony in which speaker identification efforts were made three years after the commission of the crime, and only a few words were presented in the voice sample. Yarmey (2007) found that participants who heard an unfamiliar voice once over a short duration of

time produced recognition rates below 50 percent.

Manzanero and Barón (2017) investigated the recognition accuracy rates of target-present and target-absent lineups. Participants heard a short voice sample (under 2 seconds) of 12 male voices and 12 female voices. Immediately after hearing the voices, participants were given a target-present or target-absent lineup of five voices each and ask if the original voice was present or absent in the lineup. Participants were able to identify the target voice at a rate of 83.11% but also incorrectly identified voice samples in the target-absent condition at a rate above 50%. Participants in the target-absent lineup were not given prior notification that the target-voice may not be present in the lineup. With the addition of a brief retention interval, recognition for male and female voices was below chance in target-present lineups, and false alarms were 60% for male voices and 80% for female voices in the target-absent lineups. Overall, the ability to recognize male and female target voices when tested immediately after exposure was better than when tested after a brief retention interval. In the target-absent lineups, female participants chose a female foil voice 100% of the time, suggesting a gender-bias. In a real-life context, these results show that there is a likelihood that witnesses will select a foil or another person when the perpetrator is actually in the lineup. However, the possibility that a witness may select an innocent person grows exponentially when the real perpetrator is not in the lineup.

Experiments 1 and 2 and Pilots 1 and 2 showed that participants had higher hit rates for male voices than female voices. Finding the trend that listeners identified male voices better than female voices is interesting because in real-life situations men are statistically more likely to be the primary perpetrator (Prisonstudies.org, 2019b) and victims of both genders will likely have to identify a male perpetrator more often than a female perpetrator. However, in the aforementioned experiments, false alarm

rates were much higher for male voices than female voices. If listeners have more difficulty identifying male voices with a high-level of accuracy than female voices, it is essential to consider this in applied settings.

Research has shown that women outperform men in facial recognition (Rehman & Herlitz, 2007, as cited in Areh, 2011) and show a gender-bias in visual identification of faces as well (Wright & Sladden, 2003, as cited in Areh, 2011). However, evidence of gender-bias has been conflicting and further exploration is necessary to truly determine how often it occurs and under what circumstances it is likely to occur. Varying eyewitness and earwitness strategies still have not provided sufficient support that a gender-bias exists for audiovisual or auditory recognition. This is unfortunate given the current climate of the penal systems in the U.S. and the UK where the male prisoner population exceeds the female prisoner population. This gender discrepancy will present more obstacles to accurate eyewitness and earwitness identification if the boundary conditions for it are not fully understood.

The memory processes behind eyewitness identification include encoding, storage, and retrieval. How the eyewitness encodes the sensory information at the time of the event impacts whether that particular information will become stored and later retrieved to make a successful identification. The *encoding specificity principle* suggests that the processes that take place during the encoding stage or the initial exposure should also provide the same or similar retrieval cues that will allow the witness to match those cues to the initial encoding information to successfully retrieve that information.

Memory for specific details is critical in earwitness identification. Witnesses may be called upon to answer questions about the crime scene and the perpetrator(s). During stressful events, it is likely that capturing important details is very challenging.

The witness must attend to those details to encode them, store them, and, later retrieve them. Attention is subjective, as it varies from witness to witness (Yarmey, 1995). The degree of attention is not a quantifiable number; it is a subjective calculation based on the perception of the witness or the researcher in a lab. The presumption is that the witness has some ability to see or hear the perpetrator with some level of clarity in order to provide evidence of the crime. The level of attention is best examined by the number of voice samples presented. Arguably, past research has shown that fewer voice samples increase identification accuracy, but the specific number of voice samples varies (Goldinger, 1996). We used a total of ten voice samples, five male voices, and five female voices (each speaking four words for a total of 40 presented words). This number of samples is more than the nine samples required by the UK (Home Office, 2003) and eight suggested by Wells et al. (2020) but much less than the 20-25 samples that Hollien (2012) suggests is best. It is possible that presenting ten voice samples may have impacted identification accuracy. More voice samples may likely increase accuracy, especially when considering the short voice sample duration. Further examination of the number of voice samples will be discussed in subsequent chapters.

Chapter 4 – Experiment 3

4.1 Introduction

In an effort to reduce voice identification inaccuracies, researchers have explored the effect of exposure to longer voice sample durations. Results have shown that durations of 60 -70 seconds lead to more accurate identifications during target-present lineups (Kerstholt, Jansen, Van Amersvoort, & Broeders, 2004). The UK requires at least one minute for voice samples presented in a lineup (Home Office, 2003). In the U.S., Hollien (2012) recommends that law enforce include 1- to 2-minute voice samples in their lineups. Typically, voice samples of 36 seconds or shorter have a lower accuracy rate than voice samples of 2 minutes or longer. However, higher false alarm rates are related to longer voice samples (Yarmey & Matthys, 1992). Conflicts still exist within the literature regarding the optimal length of voice samples for improved identification. If longer sample durations produce more false alarms, then very little can be done to specify an ideal length of exposure. Of course, in real-life events, witnesses may not have the benefit of experiencing lengthy exposures to voices that they will later need to recall for identification purposes or testimony.

Although voice samples of a longer duration may improve hit rates, whether they improve the recollection of content remains unresolved. In most instances, only the identification of a voice is imperative, but in instances where what was said is equally important, memory for content is essential. This area has been overlooked by past research and warrants exploration. Although the effect of voice sample duration on identification accuracy has been explored extensively, it has not been explored in relation to memory for content. Therefore, this chapter's experiment will further examine the accuracy of recognizing previously heard words and conversations.

In criminal conversations, sometimes what is said determines how well it is

remembered (Öhman et al., 2013). For example, conversations including obscene content heard over the telephone were more accurately remembered by adults than children. For adult witnesses it is more likely that stimulating conversation content like sexual or violent details will be recalled more than neutral conversation content (Pezdek & Prull, 1993). Verbatim memory is much more challenging for witnesses as they rarely recall a conversation in exact detail, but they remember the gist, or overall context of the conversation (Neisser, 1981).

Campos and Alonso-Quecuty (2006) tested the recall accuracy of a criminal conversation that participants either watched on video or heard as audio. Participants were better able to recall the gist of the conversation than verbatim details. However, participants who had to immediately recall the conversation verbatim performed better in the audio-only condition than the audiovisual condition ($M = 0.80$, $SD = 1.36$ and $M = 0.35$, $SD = 0.58$, respectively) whereby their gist recall was the same for the audiovisual condition ($M = 14.90$, $SD = 6.15$) and the audio-only condition ($M = 14.90$, $SD = 7.15$).

In Chapter 1, we mentioned that the first notable case of earwitness identification involved Charles Lindbergh's testimony against Hauptmann. Three years after a ransom drop implicating Hauptmann, Lindberg identified Hauptmann's voice as the man who demanded the ransom (*State v. Hauptmann*, 1935). Hauptmann had a German accent and it was likely that hearing the ransom suspect, who also spoke with a German accent, lead Lindberg to later select Hauptmann in a police lineup. Like the Scottish accent mentioned in the Nealon case (*R v Nealon*, 2014), voice features can impact speaker identification and lead to errors that implicate an innocent person. The idea that someone can "get it wrong" should factor into the identification process to prevent the possibility of any further evidence being tainted or

misconstrued. It is not enough to offer standardized legal criteria for court testimony after an identification has been made. In the present thesis, efforts were made to select voice samples with a neutral South East England accent. In this chapter, all speaking voices reflected the national accents representative of the UK.

Likewise, the interview recording quality may impact how well witnesses can recognize content. When analyzing a witness's exposure to a voice it is necessary to examine the presentation quality of that exposure. Typically in laboratory experiments, witnesses are exposed to voice samples through a recording device. Öhman, Eriksson, and Granhag (2010) suggest that the presentation quality can impact voice identification accuracy. They reviewed how well participants identify a speaker's voice from a voice lineup when they were previously exposed to that voice by either a recording device or a mobile phone. They found that correct identification was lower with a recording device than with a mobile phone. They did not find any significant effects of gender.

The aim of Experiment 3 was to examine how well listeners identify phrases presented in the same voices they previously heard and how well they identify voices speaking the same phrases that they heard previously. Within this context, we examined whether accuracy changed based on the speaker's gender or when the listener was presented with a new phrase. We created new voice samples of 17 to 30 seconds in duration, which are longer than the 2 second voice samples used in Experiments 1 and 2 and the pilot studies. We predicted that hearing an extended voice sample would increase the participants' accuracy for both speaker identification (H1) and content recognition (H2). We also predicted that participants would be more likely identify female speakers more accurately than male speakers (H3).

4.1.1 Method

Design

Speaker Identification condition

The experiment was a repeated measures design. Listeners heard 20 voices and the response of selecting one of two speakers was the dependent variable. The independent variable tested was the gender of the speaker (male or female).

Content Recognition condition – long duration voice samples

The experiment was a repeated measures design. Listeners heard 60 phrases spoken in various voices. The response of selecting “old” or “new” from two presented phrases was the dependent variable and the independent variable tested was the gender of the speaker (male or female).

Content Recognition condition – long and short duration voice samples

The experiment was a mixed model design. The response of selecting “old” or “new” from presented words or phrases was the dependent variable and the independent variables tested were the gender of the speaker (male or female) and the duration of the voice sample (long or short).

Participants

Twenty-nine undergraduate students at the City, University of London and non-City affiliated adults (23 females and 6 males aged between 18-44, $M = 22.72$, $SD = 7.928$) participated in the study in exchange for departmental credit or monetary compensation. All first-year Psychology undergraduates received departmental credit. Second- and third-year undergraduates, and adults recruited through the SONA Online participant management database received £4. All participants were fluent English speakers, and none reported any hearing impairments that would have

prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

Speech recordings were edited from full-length interviews on BBC's Desert Island Discs program. Interview speeches were selected due to neutral content. All speaking voices reflected national accents representative of the UK, including native British English speakers or non-native speakers with significant exposure to British English. Recordings were condensed into 17 to 30 second audio voice segments using Audacity audio software. Each segment was a separate audio file that was uploaded into E-Prime computer software and presented on a PC computer terminal. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 dB.

Procedure

The experiment was divided into five parts; a learning session, followed by a forty-five second visual filler task, a speaker identification test, followed by a forty-five second visual filler task, and content recognition test. The entire procedure took approximately 30 minutes. Before beginning the experiment, participants read an information sheet that explained their rights as a participant and signed a consent form.

In the learning session, participants were presented with twenty audio clips ranging from seventeen to thirty seconds in random order. During the presentation of each audio sample, the name of the speaker was displayed on the screen. Ten male speakers and ten female speakers spoke one phrase each. For the speaker identification test, participants were instructed that they would listen to twenty audio phrase samples spoken in various voices and after each sample, they would be asked to select the

name of the speaker from two names presented on screen. Participants were not given prior notice of the identification test before the start of the test. During the identification test, twenty audio phrases were presented individually. The voices were the same twenty voices that were originally presented in the learning session; however, new audio phrases were presented for each voice.

For the content recognition test, participants were instructed that they would listen to twenty audio phrase samples spoken in various voices. The voices were the same twenty voices that were originally presented in the learning session and speaker identification test. Ten of the audio samples included content originally presented in either the learning session or the identification test, while the remaining ten audio samples were new content. After each voice sample, they were asked to indicate if the phrase was originally spoken in either the learning session or the speaker identification test or if it was a new audio phrase. They were not given prior notice of the recognition test before the start of the test. After completing the recognition test, the participants were thanked and debriefed about the aim of the experiment.

The hit rate and correction scores were calculated based on the total response scores divided by a total number of voice samples. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1. Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding “old” and a positive score indicated the likelihood of responding “new.”

4.1.2 Results

Response Scores – Speaker Identification

All participants correctly identified more male speakers (89%) than female speakers (76%). False alarms were not calculated as the participants could only choose one of two speakers presented on screen which resulted in either a hit or a miss.

Identification Accuracy Scores

A one-way ANOVA with speaker gender (male hit rates, female hit rates) as within-subjects factor showed a significant effect of speaker gender $F(1,28) = 13.954$, $p = .001$, $\eta_p^2 = .333$.

Response scores - Content Recognition

We found that participants had slightly higher hit rates for phrases spoken in female voices (85%) than male voices (82%). Conversely, false alarms for phrases spoken in male voices were slightly lower (11%) than female voices (15%).

Table 4-1

Mean d' scores for Statement Content Recognition

	Male voices	Female voices
Mean	2.948	2.945
SD	1.204	1.401
SE	0.224	0.260
c response bias	0.008	0.201

Content Recognition – long duration voice samples

A one-way ANOVA with content (male speaker d' prime score, female speaker d' prime score) as the within-subjects factor did not show any effects of content $F(1,28) = 0.000$, $p = .992$, $\eta_p^2 = .000$.

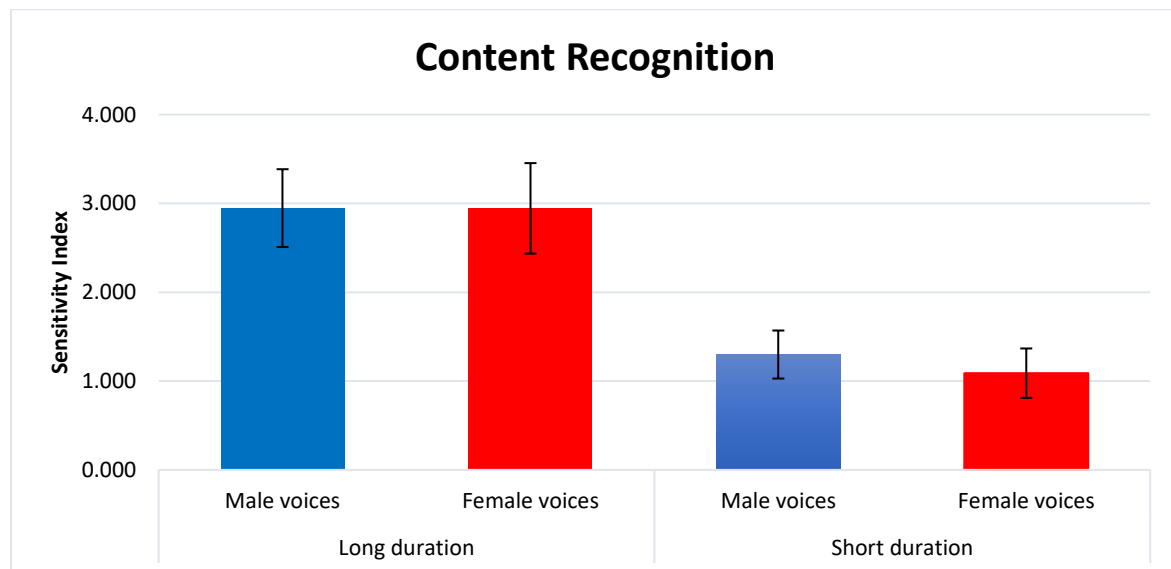
The response bias, c , for phrases spoken in male voices ranged from -1.002 to 1.002 ($M = 0.008$, $SD = 0.734$) and for female voices ranged from -2.274 to 1.255 ($M = 0.201$, $SD = 0.616$) which showed that participants were more likely to respond “new” to phrases spoken in both male and female voices.

Content Recognition – long and short duration voice samples

A 2X2 mixed model ANOVA cross-analyzed the voice samples from the word condition in Experiment 1 with the longer duration samples presented in this experiment. The voice sample duration was the between-subjects factor (long, short) and content (male speaker d' prime score, female speaker d' prime score) was the within-subjects factor. The results did not show a significant main effect of content $F(1,45) = 0.197, p = .659, \eta_p^2 = .004$ nor a significant interaction between content and duration $F(1,45) = 0.185, p = .669, \eta_p^2 = .004$. There was a significant main effect on duration $F(1,45) = 33.863, p < .001, \eta_p^2 = .429$. A univariate analysis was conducted to determine the simple main effects of duration did not yield significant results for long duration, $F(1,28) = 0.000, p = .992, \eta_p^2 = .000$ or short duration, $F(1,17) = 0.485, p = .496, \eta_p^2 = .028$.

Figure 4-1

Mean d' scores for content recognition by speaker gender as a function of duration (error bars represent confidence intervals at 95%)



4.1.3 Discussion

The results showed that participants had higher hit rates for male voices than female voices in speaker identification. A one-way ANOVA test found a significant

effect of speaker gender hit rates. Considering d' prime sensitivity was not calculated, this result does not reflect a conclusive finding nor lend support to our hypothesis.

As we read earlier, sensory memory fades. The voice sample durations in this experiment were increased from a few seconds in Experiments 1 and 2, to an average of 20 seconds to promote encoding. Additionally, the participants heard all twenty voices three times by the conclusion of the experiment. Continuous exposure to the same voice made lead to a stronger verbal memory. Participants had higher hit rates for interview content spoken in female voices than male voices while the false alarms were quite low for both genders; however, a one-way ANOVA conducted on speaker gender showed the result was non-significant.

Going further to analyze the hit rates for longer duration voice samples compared to the shorter voice samples in Experiment 1 showed higher hit rates for words and phrases spoken in male voices than female voices and less false alarms for long rather than short duration voice samples. The ANOVA result for the duration was statistically significant and suggested that voice sample length impacts witness accuracy performance, supporting our hypothesis.

According to Deffenbacher et al. (1989), the opportunity to listen is evaluated by the length of exposure to the stimulus. Bricker and Pruzansky (as cited in Yarmey, 2012) found that voice samples of at least two seconds lead to above chance identification of unfamiliar speakers. However, longer duration time improved overall accuracy performance. This suggests that current lineup procedures in the UK, where voice samples are required to be at least one minute, may produce a more accurate identification performance (Home Office, 2003). Comparatively, reducing the duration or exposure time, decreased accuracy rates (Cook and Wilding, 1997).

Although attempts were made to select neutral tone voice samples, some

variability in voice characteristics that are likely to occur during the interview process may have contributed to a limitation in the encoding process. Changes in speaking rate, pitch, amplitude, and other attributes can vary as the speaker progresses through the interview. As we previously discussed, memory encoding for voices entails an auditory signature that is compared to other voices that are subsequently presented for identification (Bradlow et al., 1999). These changes can impact attempts to accurately recognize speaking voices that were previously heard. Speaker variability threatens recognition accuracy in real-life situations where witnesses may have heard the speaker in a stressful and emotional event and later try to recognize the voice when it is unaroused.

It is also likely that speakers engaged in neutral conversation were less likely to be remembered than speakers discussing distinctive content like violent or sexual content (Pezdek & Prull, 1993). Events that include acts of physical violence resulted in a higher amount of detailed information recalled from that event than events that were non-violent (Pajón & Walsh, 2017). It is possible that some crime events are likely to include an element of violence, but it is discouraging that events involving neutral information may be less accurately remembered for recognition or recall later in time. These results could be due to the small sample size but more exploration is needed.

Chapter 5 – Experiment 4

5.1 Introduction

Manipulation of information can contribute to an interference in memory known as source confusion. A witness may experience memory interference that modifies the sensory information that they previously encoded. The exposure to interfering information like post-event manipulations creates a new memory that replaces the initial information (Smith & Baguley, 2014). Source confusion impacts identification accuracy because the witness is unable to attribute their memory to the initial source (Johnson, 1997). A witness's encoded memory is confused with a different source and the attributes of that source are retrieved during the identification process.

Post-event information can change the way a witness may recall an event. It has been suggested that witnesses tend to repeatedly recall traumatic aspects of an event and the continuous repetition leads to a more accurate recall of the central details of the event (Chan, Paterson, & van Golde, 2019). However, when witnesses observe an event but are subjected to misinformation about that event, the later recall of their observation tends to reflect the new information (Loftus & Palmer, 1974). Typically in a misinformation paradigm, witnesses are subjected to an event and are asked some leading questions that introduce misinformation about the event (Mori & Kishikawa, 2014). When witnesses are later questioned about the event, changes in their memory reports often include some of the misleading information they were exposed to after the event (Zaragoza et al., 2013).

Loftus and Palmer (1974) showed participants a collision between cars. They were asked how fast the cars were going with they collided, bumped, contacted, hit, or smashed each other. Participants gave varying estimates of the cars' speeds based

on the description of the crash. Later, participants were asked if they saw broken glass in the car crash video. Participants who were told the cars smashed into each other were more likely to report seeing glass in the video than the other participants although there was no broken glass in the video.

In Experiment 3, neutral content material was tested to determine how well listeners remembered previously heard content and if the duration of the content sample impacted memory accuracy. Building on this in Experiment 4, the content material was changed to information on actual crime scenes and events. Past research has shown that emotionally arousing material led to a more accurate recall of central details of the event and enhanced source memory (Dutton & Carroll, 2001).

The aim of Experiment 4 was to examine how well listeners identified statements they previously read (written presentation) or heard (auditory presentation). Within this context, we examined whether accuracy changed based on the speaker's gender (auditory presentation only) or when the listener was presented with an altered statement (both presentations). We changed the original details related to the crime in the written and audio statements to see if we could produce a source confusion effect. Due to the provocative content of the statements, we predicted that participants would be able to accurately recognize the original statements presented in the learning phase (H1). We predicted that participants would perform better on auditory statements than written statements (H2) and would more accurately identify statements spoken in female voices than male voices (H3).

5.1.1 Method

Design

The experiment was a mixed model design. The independent variable was the crime scenario presentation in written or auditory presentation format and the

participant's response of "Old" or "Altered" was the dependent variable. In the auditory condition, we further analyzed speaker gender as a within-subjects design and tested the effect of gender on speaker identification accuracy with the speaker's gender as the independent variable and the old/altered response as the dependent variable.

Participants

Forty adults and undergraduate students (25 females and 15 males aged between 18-56 ($M = 23.525$, $SD = 9.086$) at the City, University of London, participated in the thirty-minute study in exchange for departmental credit or monetary compensation. All first-year City, University of London Psychology undergraduates received departmental credit. Second- and third-year undergraduates, and adults recruited through City, University of London's online participant management database received £4. All participants were fluent English speakers, and none reported any hearing impairments that would have prevented participation in the experiment. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

The statements were excerpts edited from full length episodes of the television program 'Forensic Files' (see Appendix B). In both written and auditory statements, segment clips presented details of a crime that was committed or a crime scene. The statements included names of the witnesses, victims, and/or perpetrators and important dates, and locations relevant to the crime.

The written condition consisted of 20 typed-written statements that were presented on-screen to the participant. The auditory condition consisted of 18 statements presented in the written condition (9 male voices and 9 female voices). All

the statements for the auditory condition were entered into the Narrator's Voice text-to-speech mobile application on the Android platform in the Google Play Store. The speaking voices were representative of the UK including native British English speakers and selected based on their neutral accents. Due to a lack of neutral accented narrations, the researcher was not able to select an additional 10th voice for each gender. Recording files were condensed into 17 to 30 second audio samples. Each phrase was a separate audio file that was uploaded into E-Prime computer software and presented on a computer. Participants listened to each audio file on headphones with a frequency response of 20Hz – 20KHz for optimal sound quality at a volume of 10 dB.

Procedure

Of the 40 participants, 21 participants were randomly assigned to the written statement condition and 19 were assigned to the auditory statement condition. The thirty-minute experiment was divided into three parts, involving a learning session, a forty-five-second visual filler task, and a content recognition test session. In the learning session, participants were presented with twenty written statements or eighteen audio clips ranging from seventeen to thirty seconds in length.

Before beginning the experiment, participants read an information sheet that explained their rights as an experiment participant and signed a consent form. In the learning session, participants read twenty written statements presented on the computer screen or listened to eighteen audio clips played in succession. In the auditory session, nine male speakers and nine female speakers spoke one phrase each and the audio clips were presented in random order for each participant.

Before the recognition test, participants were instructed that they would read twenty statements or listen to eighteen audio phrase samples spoken in various voices.

The audio samples consisted of the eighteen voices originally presented in the learning session, but nine old and nine altered auditory statements were presented for each original voice. After each written statement or audio clip, they were required to indicate whether the statement was “OLD” or “ALTERED” from the previous statement they read or heard in the learning phrase. They were not given prior notice of the recognition test before the start of the recognition test. After completion of the recognition test, the participant was thanked for his or her participation and debriefed.

5.1.2 Results

For the written and altered statements, the participants’ responses were tallied as 0 or 1 based on signal detection measures (Stanislaw & Todorov, 1999). The hit rate and correction scores were calculated based on the total response scores divided by a total number of statements. The correct rejection score was converted to a false alarm score by subtracting the individual tallied response from 1. Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding “old” and a positive score indicated the likelihood of responding “altered.”

Response Scores

Participants had higher hit rates in the written condition (78%) than the auditory condition (74%) and false alarms were higher in the auditory condition (25%) than the written condition (18%). Participants in the auditory condition performed better with statements spoken in female voices (79%) than male voices (68%) but the false alarms were nearly equal (25% and 24%, respectively).

The response bias, c , for written statements ranged from -0.488 to 1.002 ($M = 0.112$, $SD = 0.402$) and for auditory statements ranged from -0.751 to 0.969 ($M = 0.041$, $SD = 0.441$) which showed that participants were more likely to respond “altered” to written and auditory statements.

Table 5-1

Mean d' scores for Written and Auditory Statement Recognition

	Written Statements	Auditory Statements
Mean	2.064	1.564
SD	1.101	0.609
SE	0.240	0.140
c response bias	0.112	0.041

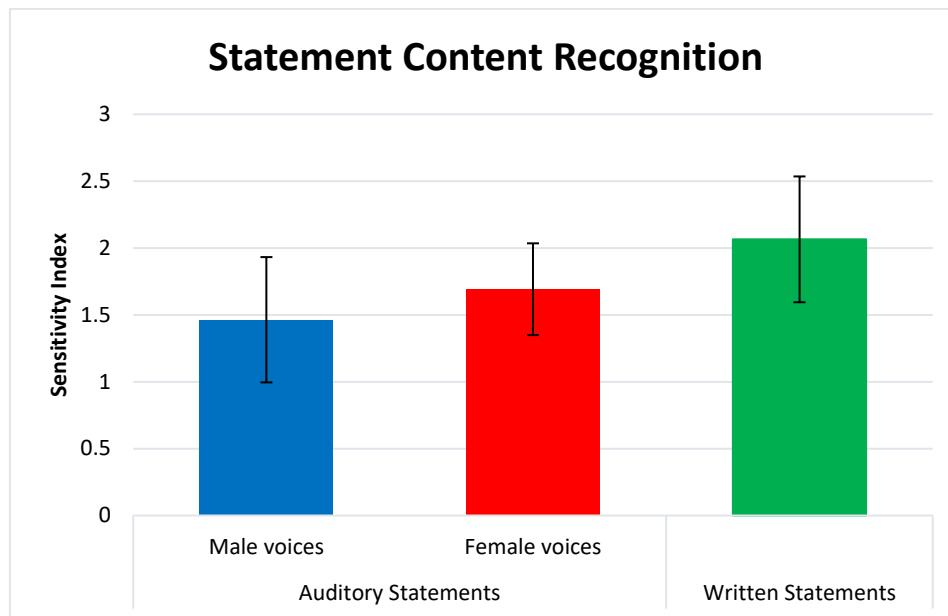
Recognition Accuracy Scores

A 2X2, mixed model ANOVA was conducted on presentation (written statement, auditory statement) as a between-subject factor and the participant’s response accuracy (old d' prime score, altered d' prime score). The results found a non-significant effect for presentation $F(1,38) = 3.073$, $p = .088$, $\eta_p^2 = .75$.

To determine if speaker gender affected response accuracy in the auditory condition, a one-way, repeated measures ANOVA was conducted on speaker gender (male, female) and response accuracy (old d' prime score, altered d' prime score). The results were non-significant for speaker gender, $F(1,18) = 0.616$, $p = .443$, $\eta_p^2 = .033$.

Figure 5-1

Mean d' scores for written and auditory statements recognition for speaker gender (error bars represent confidence intervals at 95%)



5.1.3 Discussion

The purpose of our study was to examine whether participants could detect changes in statements that they previously read or heard. We also explored if the speaker's gender impacted response accuracy in auditory statements. Based on the hit rates for both written and auditory statements, participants correctly identified the statement they previously read or heard with minimal false alarms. This trend shows that changes to the information did not affect the participants' ability to discriminate between the original and altered statements. Presenting the statements in type-written form or in audio clips did not affect performance, but analysis of the presentation mode and response accuracy did not yield a significant result. We also reviewed the impact of the speaker's gender on response accuracy and found it was not statistically significant. The gender of the speaker in the audio clips did not alter the participants' performance. Although our hypotheses were not supported, further exploration and a

larger sample population may show an effect of presentation type and speaker gender on response accuracy.

The high hit rate data trend in the auditory condition suggests that using the same voices facilitated encoding of the original statement and, thereby, made altered statements much easier to detect. In the auditory condition, the same voices were used for both the learning phase and the recognition test. Similarly, when conducting voice parades and lineups, it is important to generate a voice sample that is very similar to the voice heard by the witness during the initial event (Sherrin, 2015). The *encoding specificity principle* states that content is more likely to be remembered when it presented in the same manner during the recognition test as it was initially presented (Tulving & Thomson, 1973).

Unlike in Experiment 3, the content presented in Experiment 4 was not neutral. It contained information based on real-life criminal events that detailed the appearance and behavior of the perpetrators and victims and provided specific details about the event itself. Information including the victim's name, the time of the criminal event, and other pertinent details, created a short story of the event. Due to the provocative nature of the written and auditory content, the data trend suggested that participants were able to remember more of the original content and, therefore, able to acknowledge stark differences when tested. Previous research has shown that more explicit content is remembered with more accuracy than neutral information (Pezdek & Prull, 1993). Likewise, central details of emotionally arousing events are correctly recalled at a higher rate and with less source confusion than events that are low in emotional arousal (Dutton & Carroll, 2001).

In our study, the retention interval was short. A longer retention interval may contribute to memory misinformation (Smith & Baguley, 2014). The recognition test

was conducted immediately after the participants completed a short visual filled task. The reduction of time between the initial exposure and the recognition test would be less likely to impact any encoding of distorted information that may lead to poorer recognition of altered statements (Smith & Baguley, 2014). Although we did not test the effects of the retention interval, it could be suggested that conducting the recognition test shortly after the learning phase could have contributed to the participants ability to distinguish original statements from altered statements with higher accuracy.

The trend effects of the analysis suggest that participants accurately recognized original statements better than altered statements in both written and auditory presentations. Participants also recognized statements spoken in a female voice better than those spoken in a male voice; however, any gender differences that were presented should not be taken into account as the sample size of nine voices for each gender was small. The exploration of speaker gender differences should only be acknowledged as a prospective distinction but not a definitive difference among the general population. This experiment did not evaluate speaker identification; therefore, gender-bias detection was not addressed.

Chapter 6 – Experiment 5

6.1 Introduction

One of the challenges earwitnesses face is being called upon to recognize the unfamiliar voice of a perpetrator. Most witnesses are exposed to voices that they have heard in the first instance for only a brief duration of time. One may assume that a witness can easily identify a voice that is considered familiar without a voice lineup because they already know who the speaker is. This idea is based on the notion that a listener's level of familiarity with a speaker is always high. However, one can be familiar with a voice in varying ways. Hollien (2012) proposed different levels of familiarity and distinguished between “just barely familiar,” “kind of familiar” and “very familiar.” He found that witnesses can recognize voices that are “very familiar” much better than those that are less familiar. Similarly, Yarmey (2007) found that witnesses were more accurate at recognizing “highly familiar” voices than “moderately familiar” and “not-so-familiar voices (85%, 79%, and 49%, respectively). However, there is very little research that examines levels of familiarity. The extent of the research has examined familiarity in a general sense as it may relate to family members (Yarmey, 2012) or TV characters (Lavan et al., 2019). Therefore, it is difficult to determine whether previous results on familiarity are applicable when target voices are only considered “kind of familiar” or “just barely familiar” by the witness.

Furthermore, there is no guarantee that a witness will accurately recognize a familiar voice (Read & Craik, 1995). Changes in distinctive characteristics may make the identification of a familiar voice challenging (Read & Craik, 1995; Yarmey, 1995, 2012). The length of a voice sample can impact identification for familiar and unfamiliar voices. Extending an utterance from a single word to a 30-second phrase

increases identification accuracy from 31% to 83%, even when identifying a mother's voice (Ladefoged & Ladefoged, 1980).

Considering confidence is also misleading because a higher confidence level does not suggest a higher level of identification accuracy. Research has shown a stronger correlation between confidence and accuracy when identifying very familiar voices than unfamiliar voices (Yarmey et al., 2001). Sarwar, Allwood, and Zetterholm (2014) analyzed whether there was an overall relationship between higher confidence ratings and identification accuracy. The participants listened to a dialogue between two unfamiliar perpetrators planning a burglary. One man was the leader and spoke over 70% of the time, while another man was an accomplice. After a 15-minute retention interval, the participants were asked to identify the voice of the main speaker in a voice lineup that consisted of either a text lineup or a dialogue lineup. In the text lineup, the target speaker and five foils read the same text. In the dialogue lineup, the target speaker and five foils discussed a newspaper article with an inaudible companion. Results showed that overall, only 37% of participants in both conditions selected the target speaker within a group, and their mean confidence level was 64.41% ($SD = 21.27$). The remaining participants either selected one of the foil voices or did not recognize any of the voices but their confidence levels exceeded 50% ($M = 53.86\%$, $SD = 23.53$ and $M = 54.36\%$, $SD = 28.73$, respectively). The difference in mean confidence levels between correct and incorrect identification was only around 10%. It is alarming that witnesses can confidently support their identification or lack thereof when it is inaccurate. The worry is that jurors will assign a significant amount of weight to confidence levels when there lacks certainty that it adequately reflects accuracy.

In the previous studies, we examined how well people identified voices and

whether that varied depending on the speaker's gender or when presented with a new voice. This study builds on this by analyzing whether familiarity of a voice can predict how well people remember voices and whether confidence is a predictor of speaker identification accuracy. In the present study, voice familiarity was measured on a 5-point Likert scale to determine levels similar to those reported by Yarmey (2007), which ranged from "extremely familiar" to "not familiar at all." We also requested that participants provide a confidence rating on the accuracy of their identification choice. The confidence rating was measured on a 7-point Likert scale ranging from "Not very confident" to "Very confident." We predicted that participants would more accurately identify familiar voices than unfamiliar voices (H1) and they would give higher confidence scores for correct identifications (H2).

6.1.1 Method

Design

The study is a within-subjects design to test whether there is relationship between voice familiarity, speaker identification accuracy, and confidence ratings. Familiarity and confidence ratings were the predictor variables and speaker identification accuracy was the outcome variable.

Participants

Thirty-two adults (20 females and 12 males aged between 22-58, $M = 33.78$, $SD = 9.51$) recruited through Amazon Mechanical Turk participated in the study in exchange for monetary compensation of £4. All participants were fluent English speakers from the United Kingdom, and none reported any hearing impairments that would have prevented participation in the experiment. All participants were required to use a desktop or laptop computer in a private setting for optimal testing conditions

and software performance. The City, University of London Research Ethics Committee granted approval for the experiment.

Materials

Speech recordings were edited from full-length interviews on BBC's Desert Island Discs program. The same voice samples from Experiment 3 were presented. Each phrase was a separate audio file that was uploaded into Qualtrics Survey Software.

Procedure

Each experiment was divided into three parts: 1) a learning session, 2) a ten-minute rest break, and 3) a recognition test session (see Appendix C). The same voice samples from Experiment 3 were presented. Before beginning the experiment, participants read an information sheet that explained their rights as a participant and submitted an electronic consent form.

In the learning session, participants were presented with ten audio clips ranging from seventeen to thirty seconds in duration. The clips were presented in random order. Five male speakers and five female speakers spoke one phrase each. After each phrase, participants had to rate the familiarity of the voice on a 5-point Likert scale ranging from "Extremely familiar" to "Not familiar at all." After completing the learning session, participants were given a ten-minute rest break before the speaker identification test.

Before the speaker identification test, participants were instructed to listen to twenty audio phrase samples spoken in various voices. The audio samples consisted of ten voices originally presented in the learning session and ten new voice samples. After each audio sample, participants selected the button for an "old voice" presented during the learning session or the button for a "new voice." After they indicated old

or new, they were asked to rate their confidence in the accuracy of their responses on a 7-point Likert scale ranging from “Not very confident” to “Very confident.” Participants were not given prior notice of the speaker identification test before the start of the test. After completing the speaker identification test, the participant was thanked for his or her participation and debriefed regarding the aim of the experiment.

6.1.2 Results

Scores for hit rates and false alarm rates were converted into z scores, and the d' prime score was calculated by subtracting the false alarm z score from the hit rate z score. The response bias c score was calculated by averaging the z scores for hit and false alarm rates. A negative c score indicated the likelihood of responding “old” and a positive score indicated the likelihood of responding “new.”

Response Scores

Participants had hit rates above chance (63%, $SD = 2.422$) and false alarms were moderate (34%, $SD = 0.212$). The response bias, c , for speaker identification ranged from -0.380 to 0.578 ($M = 0.042$, $SD = 0.199$) which showed that participants were more likely to respond “new” to the voice samples.

Table 6-1

Mean d' scores for Speaker Identification and mean scores for Familiarity and Confidence ratings

	d' prime	Familiarity rating	Confidence rating
Mean	0.860	2.688	4.916
SD	1.109	0.747	1.045
SE	0.196	0.132	0.184
c response bias	0.042		

Identification Accuracy Scores

A multiple regression with enter method was used to predict speaker identification from familiarity and confidence rating based on d' prime criterion. The model did not show a statistically significant amount of variance in speaker identification, $F(2,29) = 0.109, p = .897, R^2 = 0.007, R^2_{\text{adjusted}} = -0.61$. Confidence was not a significant predictor for speaker identification accuracy, $B = 0.008, t(29) = 0.041, p = .968$. An increase of one confidence rating corresponded to a slight increase in speaker identification score of 0.004 points, $B = 0.004, 95\% \text{ CI } [-0.179, 0.186]$. Familiarity was also not a significant predictor for speaker identification accuracy, $B = -0.086, t(29) = -0.464, p = .646$. An increase of one familiarity rating corresponded to a decrease in speaker identification accuracy by 0.088 points, $B = -0.088, 95\% \text{ CI } [-0.478, 0.301]$.

A multiple regression with enter method was used to predict speaker identification from familiarity and confidence rating based on c response bias criterion. The model did not show a statistically significant amount of variance in speaker identification, $F(2,29) = 0.274, p = .763, R^2 = 0.019, R^2_{\text{adjusted}} = -0.049$. Confidence was not a significant predictor for speaker identification accuracy, $B = -0.096, t(29) = -0.520, p = .607$. An increase of one confidence rating corresponded to a slight decrease in speaker identification accuracy of 0.025 points, $B = -0.025, 95\% \text{ CI } [-0.122, 0.073]$. Familiarity was also not a significant predictor for speaker identification accuracy, $B = -0.100, t(29) = -0.545, p = .590$. An increase of one familiarity rating corresponded to a decrease in speaker identification accuracy by 0.055 points, $B = -0.055, 95\% \text{ CI } [-0.264, 0.153]$.

Further review of the ROC curve for speaker identification in Figure 6-1 shows no significant discrimination between old and new voices, $AUC = 0.608$. Analyzing the ROC curves of familiarity (see Figure 6-2) and confidence ratings (see

Figure 6-3) with speaker identification accuracy did not yield significant discriminability between old and new voices ($AUC = 0.313$, $AUC = 0.351$, respectively).

Figure 6-1

ROC curve for Speaker Identification Accuracy

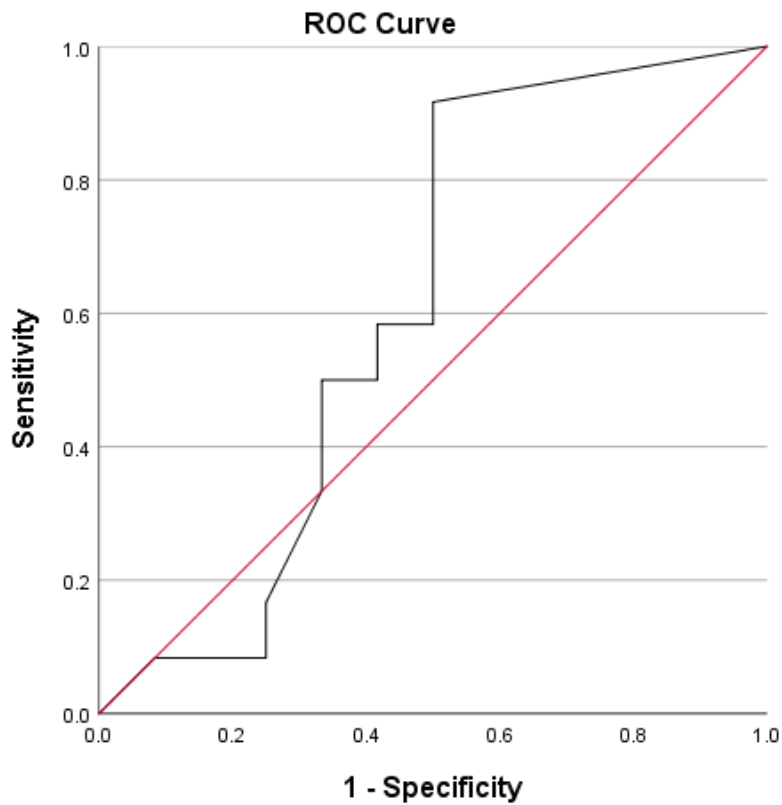


Figure 6-2

ROC curve for Speaker Identification Accuracy and Familiarity

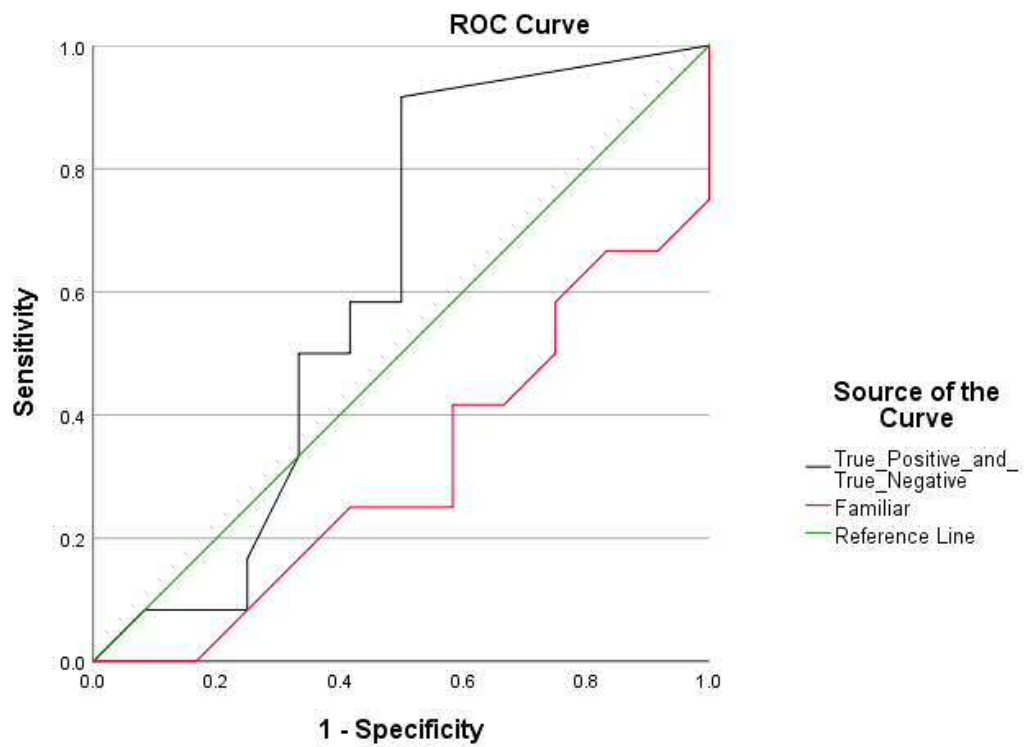
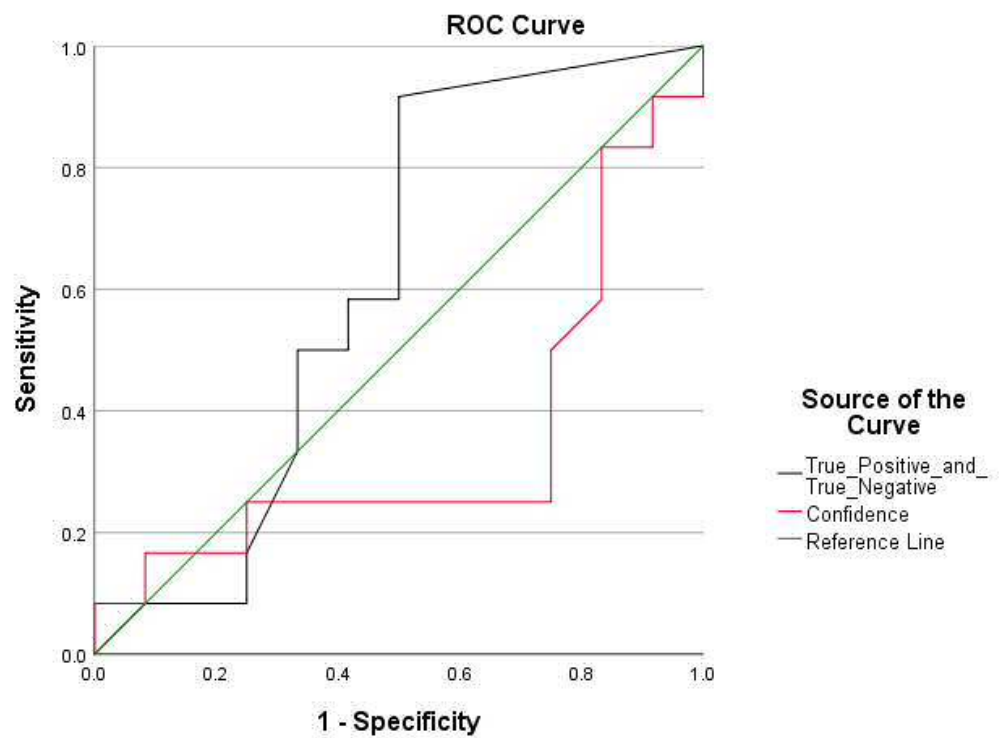


Figure 6-3

ROC curve for Speaker Identification and Confidence



6.1.3 Discussion

The data trends showed that a one point increase in the familiarity rating leads to a decrease in speaker identification accuracy for both d' prime and response bias criteria. However, a one point increase in the confidence rating leads to a slight increase for speaker identification in the d' prime criterion but a decrease for response bias criterion. Evaluating the means for the hit rates, participants identified speakers at a rate higher than chance but the results did not show that there was a statistically significant relationship between speaker identification accuracy and familiarity or confidence ratings and, therefore, the hypotheses are not supported.

According to the *encoding specificity principle*, participants would more accurately identify voice samples during a lineup that matched the initial voice samples (Tulving & Thomson, 1973). In this study, participants were presented with the same ten voices during the lineup they heard during the learning phase. Although it seems likely that presenting the same speaking voices would increase accuracy, the results did not show this to be the case. In our findings, the participants reported low familiarity ratings, with the average showing a “slight familiarity” for the voices presented in the learning phase. We did not find a significant relationship between familiarity and speaking voice accuracy. The accuracy of recognizing familiar voices can vary greatly (Yarmey, 1995). Witnesses focus on the features of unfamiliar voices rather than on the pattern, like they do for familiar voices (Yarmey et al., 2001). Voices that are unfamiliar, offer other unique traits which may be perceived as distinctive to one witness but less distinctive to another.

One factor that is highly researched in the area of eyewitness identification is confidence ratings. Substantial weight is given to an identification that is supported with high confidence levels from the eyewitness. In our findings, the participants

reported moderate to high confidence ratings for speaker identifications; however, we did not find a significant relationship between accuracy and confidence and few research studies have found that accurate speaker identification is related to confidence. Malpass and Devine (1981) explained that eyewitnesses are more focused on making an identification rather than making an “accurate” identification. This disparity is concerning and the possibility of continued reliance on confidence ratings for testimonial support would lead to suggestibility.

Witnesses can, and do, make errors despite often being convinced otherwise. Memory processes make accurately recognizing voices or recalling spoken content very challenging. In James Marcello's case, the victim's daughter identified Marcello's voice three years later as the same voice who called her father the day her father disappeared and was later found dead. The witness stated that she was “100 percent sure” of the voice in her court testimony (Saltzburg, 2013). Marcello tried, unsuccessfully, to introduce expert testimony from Daniel Yarmey arguing that voice identifications are often mistaken. Yarmey intended to introduce evidence of suggestibility based on his analysis of the Federal Bureau of Investigation's (FBI) voice lineup whereby witnesses chose Marcello's voice out of a lineup at a rate higher than chance. This evidence would have given credence to the notion that the FBI lineup was not neutral and Marcello's voice was more likely to be selected by the victim's daughter. Unfortunately, the judge did not allow the expert testimony and declared that the jury would not find any difference in the voice samples (Saltzburg, 2013). Jurors attribute a substantial amount of weight to testimony where a witness has identified a suspect and are more likely to pass down guilty verdicts in such circumstances (Loftus, 1975).

Chapter 7 – Experiment 6

7.1 Introduction

In the previous chapters, we have reviewed various factors that influence eyewitness identification. Here, we proposed a technological mnemonic that corroborates evidence rather than relying on human memory. Of course, this is merely speculative as no evidence shows any impact of mobile applications on identification accuracy. The hope is that further digital developments may help to reduce identification errors.

As previous research has shown, memory is delicate and can be fallible at times. Witnesses are called upon to contribute evidence of what they remember as part of a criminal investigation. Police departments receive criminal reports through several mediums, ranging from over-the-phone contact, online website reporting, in-person written reports, and more. There are many ways that information can be routed to the correct authorities and many ways that problems can arise en route (Vatanasuk, Chomputawat, Chomputawat, & Chatwiriya, 2015). For example, researchers in Thailand identified four issues that Bangkok police departments struggle with when using conventional reporting methods: (1) insufficient details to aid in response or support other enforcement agencies, (2) misleading or fabricated reports, (3) poorly trained response teams, and (4) inadequate data collection measures to analyze crime statistics (Vatanasuk et al., 2015).

To resolve the issues faced in Bangkok and other parts of the world, the researchers proposed a mobile application that will collect all the relevant information into a single report that could be sent to a response team (Vatanasuk et al., 2015). The structure of their mobile application is similar to the Self Evidence application that was available in the UK (The Smart Way to Report Crime, 2019). Self Evident

addressed several issues that law enforcement agencies face when dealing with crime reports and responding to incidents. The application controlled for fraudulent reports by requiring registration and user validation. It also collected substantive details of the incident and assisted the user in sending the report to the proper authorities, thus reducing response team error. Lastly, it collected all the data in an incident report that could be cataloged for future reference and analysis. The idea of using a mobile application for crime reporting is not a novel one. However, the main obstacle that applications like Self Evident still face is a lack of users and, worse, a lack of support from law enforcement. The Self Evident application was launched in 2013 and had less than 30,000 users and approximately 10,000 downloads on the Google Play Store (Android operating system) before it was suspended in November 2018 due to a lack of funding for the application developer and charity, Witness Confident (BBC, 2018).

While Self Evident was on the brink of suspension in the UK, researchers at the University of Sydney in Australia launched their crime reporting mobile application, iWitnessed (Paterson, van Golde, Devery, Cowdery, & Kemp, 2018). Like Self Evident, iWitnessed is available for free to download on all smartphones; however, it is unavailable to download in the UK. The researchers suggested that as smartphone users continue to increase, the need to streamline eyewitness accounts and incident reports is imperative. As of 2018, over 84 percent of Australians have a smartphone (Deloitte, 2016 as cited in Paterson et al., 2018) and use will continue to rise globally as users get more acclimated to technological advances at a much younger age than their predecessors. Teens and millennials are now classified as the “mobile youth culture” because they have been exposed to and use mobile phones with higher frequency than previous generational cohorts (Vanden Abeele, 2016). Teens use smartphones to communicate, access information, express creativity, and

more. Nowadays, teenagers prefer to communicate with peers using messaging mobile applications like WhatsApp rather than making voice phone calls (Vanden Abeele, 2016). Today's mobile youth culture will influence future smartphone directives that will launch society into the next phase of communication innovation. The global population increasingly relies on instant hand-held access. Applications like Self Evident and iWitnessed are essential for the general public's access to a safe space for reporting incidents like hate crimes and domestic abuse.

The aim of this study was to determine how well participants can recall details of a crime. The rationale behind this study was to address the recalled details of a criminal event and assess the utility of using a mobile application to complete a crime report. The purpose was to analyze how participants engaged with a crime reporting mobile application and input recall data. At the time of the study, the mobile application prototype was not fully functional. To simulate aspects of the mobile application, I created a crime report on Qualtrics to collect recall data similar to the mobile application I intend to launch. The simulation provided insight into how people interact with technology and how accurately they input details into an electronic display system. My mobile application, *Provide the Proof* (see Appendix F for screenshot photographs), will offer features similar to Self Evident and promote personal safety through emergency assistance and notification options (in future updates). The application will log an incident and capture real-time images and voice and video recordings. It will link the report file to global positioning (GPS) and personal contact details that could be logged for future retrieval or to upload to law enforcement and legal professionals. The ease of global access at our fingertips perpetuates the idea that applications like incident reporting will be increasingly used

by the “mobile youth culture” who continue to influence how we adapt to new communication channels.

Experiment 6

To further reduce eyewitness identification inaccuracies, it is imperative to create a resource that can be used in tandem with digital technologies accessible to the public. A mobile application that captures crime information in real-time or immediately after a crime has occurred has the potential to reduce false memories and misinformation issues that are likely to occur over time. Therefore, I created an online simulation of a mobile application using Qualtrics Survey Software to capture the design's fluidity and evaluate users' responses. We predicted that participants would be able to recall idea units from a crime event more accurately when the victim was female rather than male (H1). This prediction is based on prior research which showed that men and women have a more accurate memory for the appearance, or descriptive details relating to female victims (Horgan, Mast, Hall, & Carter, 2004). We also predicted that female participants would recall more idea units than males (H2) based on previous research showing that women perform better than men when recalling the appearance of male and female targets (Horgan et al., 2004).

7.1.1 Method

Design

The experiment was a between-subjects design with four conditions: burglary, domestic violence, harassment, and motor incident. We analyzed the crime scenarios and the victim's gender as predictors and the recall accuracy as the outcome.

Participants

Thirty adults (21 males and 9 females, $M = 31.166$, $SD = 7.737$) recruited through Amazon Mechanical Turk participated in the study in exchange for monetary

compensation of £4. Out of the total participants, 9 were randomly allocated to the burglary scenario, 12 allocated to domestic violence, 3 allocated to harassment, and 6 allocated to motor incident. All participants were fluent English speakers from the United Kingdom, and none reported any hearing impairments that would have prevented participation in the experiment. All participants were required to use a desktop or laptop computer in a private setting for optimal testing conditions and software performance. The City, University of London Research Ethics Committee granted approval for the experiment.

Stimulus Material

Information was adapted from real-life crimes presented on ‘Forensic Files’ into four crime incident scenarios (see Appendix E). The scenarios were transcribed into text and the text was converted to an audio sample recording using the Narrator’s Voice text-to-speech mobile application on the Android platform in the Google Play Store. Four narrator’s voices (2 males and 2 females) were selected based on a neutral, South Eastern England accent. Each individual audio recording was uploaded into Qualtrics Survey Software.

In the recall task, participants completed a survey that asked both forced choice questions and open recall questions (see Appendix D). The forced choice questions asked the gender of the responding officer and victim in the scenario. The recall questions asked for details relating to the crime scenario such as, “What time did the incident take place?”. Participants were given an unrestricted amount of time to complete the task.

Procedure

Each experiment was divided into three parts; a learning session, a ten-minute rest break, and a recall test session. Before beginning the experiment, participants read

an information sheet that explained their rights as experimental participants and submitted an electronic consent form.

In the learning session, participants were presented with an option to self-randomize by selecting one of four colors. Each color corresponded to an audio recording of a crime scenario: Red (Domestic Violence), Blue (Burglary), Green (Motor incident), and Yellow (Harassment). After the participant selected a color, the audio recording of the crime scenario was presented. Each crime included one police officer (speaker) and one victim. Out of the four scenarios, two included male police officers and two included female police officers. The gender of the victim(s) was the opposite of the police officer. After listening to the crime scenario, participants were given a ten-minute break before continuing to the crime report recall test.

Before the recall test, participants were instructed that they would complete a crime report based on the audio recording they had heard during the learning session. In the recall test, participants had to indicate the type of crime (burglary, domestic violence, harassment, motor incident), the gender of the police officer, the gender of the victim, the date of the crime, the location where the incident occurred, and give an open-ended description of the crime. Upon completion of the survey, participants were debriefed and thanked for their time.

Statements for each crime report were coded into idea units, or nodes (Mandler & Johnson, 1977). An idea unit is essentially a phrase or sentence that contains a subject and a verb. Participants received one point for including each of the following details from the crime: time, date, and location, witness(es) gender, officer(s) gender and any further recall of important details (see Appendix E). Recall accuracy was calculated as the number of correct idea units out of the total number of idea units in each crime category. A second independent observer scored the

recall idea units for 50% of the total surveys. The differences were discussed until the observers settled upon an agreement. The interrater reliability was $r = .937$, $p < .001$.

7.1.2 Results

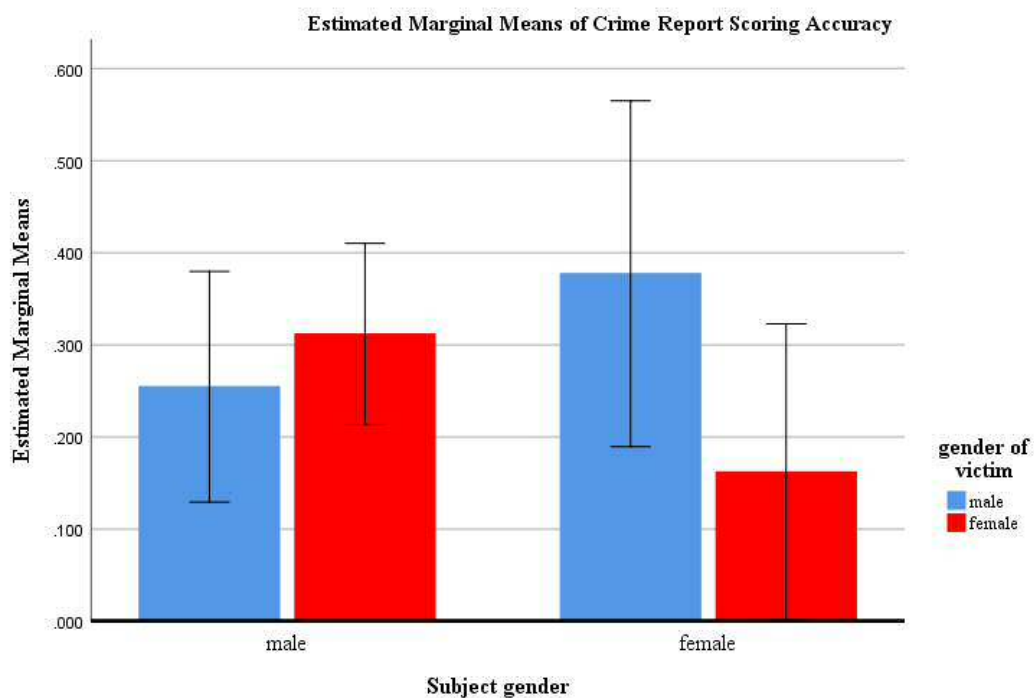
Recall accuracy was calculated by the number of correct idea units out of the total number of idea units in each crime category. The total number of idea units in each crime scenario ranged from 14 to 19 (Motor incident – 14, burglary – 15, harassment – 15, and domestic violence – 19). The final idea unit scores were analyzed based on the gender of the speaker and the description of the crime. Male participants recalled more idea units in crime scenarios where the police officer was male and the victim was female ($M = 0.315$, $SD = 0.166$) than when the police officer was female and the victim was male ($M = 0.251$, $SD = 0.173$). Conversely, female participants reported more idea units for scenarios in which the police officer was female and the victim was male ($M = 0.367$, $SD = 0.139$) than vice versa ($M = 0.169$, $SD = 0.164$). Overall, details in the crime scenarios involving male victims ($M = 0.289$, $SD = 0.166$) were recalled slightly more accurately than crimes with female victims ($M = 0.274$, $SD = 0.174$).

A 2x2 ANCOVA was conducted to compare the effects of participant gender (male, female) and victim gender (male, female) while controlling for the crime scenario (burglar, domestic violence, motor incident, harassment). This was to determine whether the victim's gender impacted the number of correct recalled idea units among the male and female participants. There were no significant differences in mean participant gender, $F(1,25) = 0.040$, $p = .844$, $\eta_p^2 = .002$ and victim gender, $F(1,25) = 0.996$, $p = .328$, $\eta_p^2 = 0.038$. There was no significant interaction between participant and victim gender, $F(1,25) = 3.820$, $p = .062$, $\eta_p^2 = .133$.

The estimated marginal means data trend showed that males performed slightly better on the recall task than females ($M = 0.283$, $M = 0.270$, respectively) and more idea units were recalled when the victim was male than female ($M = 0.316$, $M = 0.237$, respectively).

Figure 7-1

Mean recall idea unit scores for participant gender as a function of speaker gender (error bars represent confidence intervals at 95%)



Male participants recalled more idea units in domestic violence ($M = 0.336$, $SD = 0.168$) and motor incident ($M = 0.243$, $SD = 0.199$) crime scenarios than female participants ($M = 0.211$, $SD = 0.155$ and $M = 0.001$, $SD = 0.00$, respectively, as there was only one participant). The female participants ($M = 0.467$, $SD = 0.094$) recalled more idea units in the harassment crime scenario than male participants ($M = 0.333$, $SD = 0.00$, as there was only one participant). Both male and female participants recalled the same amount of idea units for the burglary scenario ($M = 0.267$, $SD =$

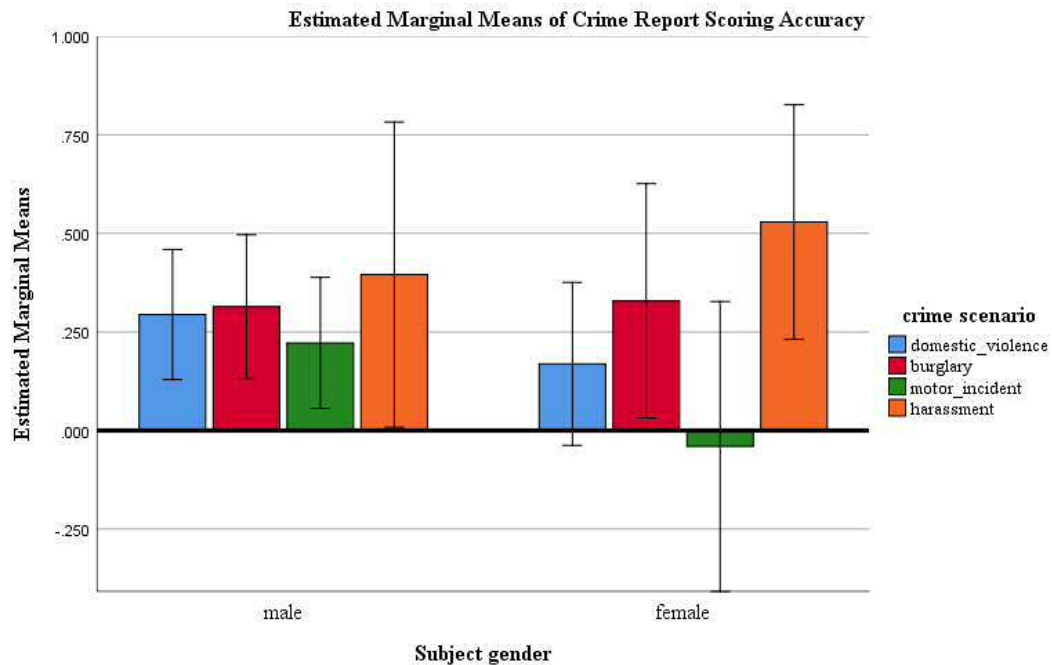
0.168 and $M = 0.267$, $SD = 0.094$, respectively). Overall, males performed slightly better on the recall task than females ($M = 0.290$, $SD = 0.167$ and $M = 0.257$, $SD = 0.177$).

A 2x4 ANCOVA was conducted to compare the effects of participant gender (male, female) and crime scenario (burglar, domestic violence, motor incident, harassment) while controlling for victim gender (male, female). This was to determine whether the type of crime scenario impacted the number of correct recalled idea units among the male and female participants. There were no significant differences in mean participant gender, $F(1,21) = 0.542$, $p = .470$, $\eta_p^2 = .025$ and crime scenario, $F(1,21) = 1.537$, $p = .234$, $\eta_p^2 = 0.180$. There was no significant interaction between participant and crime scenario, $F(1,21) = 0.888$, $p = .463$, $\eta_p^2 = .113$.

The estimated marginal means data trend showed that males performed slightly better on the recall task than females ($M = 0.306$, $M = 0.247$ respectively) and more idea units were recalled in the harassment scenario ($M = 0.462$) compared to domestic violence ($M = 0.232$), burglary ($M = 0.322$), and motor incident ($M = 0.091$), respectively.

Figure 7-2

Mean recall idea unit scores for participant gender as a function of crime scenario (error bars represent confidence intervals at 95%)



7.1.3 Discussion

The data trend showed that males correctly recalled more idea units when the victim was female and vice versa for female participants. Unlike Horgan et al. (2004), we did not find that females outperformed men in recalling crime scenario idea units. The trend also showed that male participants recalled slightly more idea units across all crime scenarios than females. The results were not statistically significant so our hypotheses were not supported. The unequal randomization of the crime scenario types among participants likely contributed to this non-significant result. During the initial design phase, the researcher was unable to properly randomize the survey and, therefore, the alternative color choice was implemented. This, however, also

prevented an equal distribution among the participants and a more effective randomization process should be conducted in the future to prevent such discrepancies.

Although the data trends did not show that participants recalled the idea units with a high level of accuracy, the written modality implemented in the study could contribute to more accurate recall than a spoken account of the crime details. Sauerland, Kirx, van Kan, Glunz, and Sak (2014) suggest that witnesses produce accounts of events in written and spoken voice modalities and both are influential in recall accuracy. Past research has leaned more toward the spoken voice as the superior modality in recall (Sauerland & Sporer, 2011). However, other studies have shown that written accounts could be more accurate than spoken accounts. Grabowski (2005) found that students could recall more European countries and capitals when they wrote them rather than speaking them. He attributes this effect to using less cognitive resources when writing rather than speaking. Fueller, Loescher, Indefrey (2013) supports this claim and found that recall was more accurate when the participants wrote their account of both visual and auditory input rather than giving a spoken account of the details. Based on this research, it is possible that a significant effect could be achieved with the written crime scenario recall but the smaller sample size in our study may likely have reduced recall accuracy.

There is no evidence to support that providing the exact words or phrases spoken by the perpetrator during a lineup will increase identification accuracy (Reisberg, 2014). However, remembering what may have been said could be helpful, especially with crimes like harassment or fraud. As I mentioned in Chapter 2, memory for content goes above and beyond just recognizing a voice (Reisberg, 2014) because

it may involve an evaluation of what was said. This does not mean a verbatim retention of content, but an exploration of the overall content and central understanding of the conversation.

In the coming years, more and more interactions with law enforcement and legal professionals will occur remotely. The utility of a crime reporting mobile application is to reduce inaccuracies by offering features that will reduce the need to recall event details in the future. The features will include a time, date, and a global positioning stamp that will be embedded within the report file so that errors attributable to memory will be reduced or eliminated. Capitalizing on previous research by Sauerland et al. (2014) and Fueller et al. (2013), the application will capture both written and audio descriptions that can be accompanied by photo imagery and videos so that descriptive details such as perpetrator characteristics, victim descriptions, and other details can be succinctly recorded and linked to the report file.

To reduce misidentification, it is important that witnesses can report events with some level of immediacy, if warranted, like in cases of assault or robbery. Further delays between witnessing an event and, possibly, participating in a lineup may lead to the witness forgetting the details of the event or becoming susceptible to misinformation (Paterson et al., 2018). Law enforcement is inundated with paperwork and telephone calls that are the current channels in place for witnesses to report incidents and crimes (Paterson et al., 2018). Offering an alternative that is accessible to most of the population with smartphones is a solution that can free up law enforcement to conduct routine police work. We continue to progress to a more digital domain that serves our daily needs and streamlining safety measures is a step forward in the same digital continuum. Mobile applications like Provide the Proof and

iWitnessed will help to transform how we collaborate with security personal and offer an alternative platform that will safeguard valuable evidence.

Chapter 8 - General Discussion

The aim of this thesis was to determine: (1) how well do witnesses remember voices, (2) does speaker identification accuracy vary with the gender of the speaker, (3) does speaker identification accuracy vary when the witness is presented with a new voice or new phrase, (4) does speaker familiarity or confidence ratings predict speaker identification accuracy, (5) how well do witnesses recall details of a crime. These questions were addressed in six experiments and two pilot studies that analyzed speaker identification and memory recall accuracy. The overarching objectives were to determine conditions that may influence speaker accuracy and impact free recall.

As previous research has shown, it is increasingly likely that auditory information will be presented as evidence during court testimony. The need for more accurate earwitness testimony is essential to offering significant evidence that could aid in the successful conviction of the correct perpetrator or at least reduce the possibility of wrongfully convicting an innocent person. However, research has suggested that memory for remembering voices is flawed, and the efforts to reduce errors are still vital. In this final chapter, I will: (1) address the main findings of each experiment, (2) discuss theoretical implications and how the findings relate to previous research, (3) address the limitations of the findings and, (4) conclude with suggestions for future research on speaker identification and free recall accuracy.

8.1 Core research findings

In Experiment 1, participants listened to twenty voice samples. Ten male and ten female speakers spoke four words each, and the participants were given a recognition test with 40 voice samples to determine if they could identify a word or voice that they previously heard. Participants did not distinguish between old words

and new words or old voices and new voices in the recognition test to a significant level. The data trend showed that participants identified words spoken in male voices more accurately and performed better with male voices than female voices overall.

In Pilots 1 and 2, changes in stimuli presentation to improve encoding proved futile. Therefore, we decided to modify the content with provocative stimuli to foster encoding. In Experiment 2, participants read a physically violent robbery crime scenario before listening to the same voice samples presented in Experiment 1. The participants completed a recognition test to determine if they could identify the voices they previously heard. There was a significant main effect of the speaker's gender. The data trend showed that similar to Experiment 1, participants identified male voices with more accuracy than female voices, but the false alarm rates were much higher for males than females.

In Experiment 3, participants listened to ten voice samples with five male and five female speakers. They were subsequently tested on whether they recognized the speaker from the initial voice sample presentation, followed by a test to determine if they remembered what the speakers discussed in the auditory interview clips. There was no statistically significant effect of the speaker's gender on recognizing neutral interview content; however, analyzing only the speaker identification hit rates showed a significant effect of the speaker's gender on identification accuracy. The participants accurately recalled more male voices than female voices, but the small sample size does not adequately suggest generalizability.

Further analysis of voice sample duration between Experiment 1 and Experiment 3 evaluated the short duration samples of ~ 2 seconds (Exp. 1) and longer duration samples of ~ 17-30 seconds (Exp. 3). The results yielded a significant main

effect for the duration but not a significant main effect of the speaker's gender or a significant interaction between the speaker's gender and duration. Further analysis did not show a simple main effect for short or long duration presentations.

In Experiment 4, we changed the type of content and the type of presentation modality. We presented a provocative crime scene from real-life crimes to participants in either a written or auditory presentation format. Participants read twenty written statements or listened to eighteen auditory statements with nine male and nine female speakers during the learning phase. Participants were given a recognition test with parts of the original statements altered to determine if they could accurately choose the original statements. Although the data trend showed higher accuracy in the written condition, the results did not show a significant effect of the presentation format on written or auditory statements. An analysis to determine the effect of the speaker's gender in the auditory format did not yield a statistically significant result; however, the data trend showed that participants identified original statements spoken in female voices at a higher rate than male voices.

In Experiment 5, we used the same stimuli presented in Experiment 3 but further examined voice familiarity and evaluated confidence ratings on speaker identification. Participants listened to ten voice samples (five males, five females) and rated their familiarity with the voice on a Likert scale of 1 (not familiar at all) to 5 (extremely familiar). Participants were given a recognition test with the original ten voices and phrases and an additional ten new voices and phrases. They were asked to determine if the spoken phrase was an old or new phrase and rate how confident they were on a Likert scale of 1 (not very confident) to 7 (very confident). We analyzed whether familiarity or confidence ratings could predict speaker identification

accuracy. We analyzed speaker identification accuracy and calculated d' prime sensitivity and response bias, and found that familiarity and confidence were not statistically significant predictors. The data trends showed that an increase of one familiarity rating point led to a slight increase in speaker identification and, conversely, led to a decrease in speaker identification with an increase in one confidence rating point. The ROC curve results comparing confidence and familiarity ratings with speaker identification did not show discriminability between old and new voices to a significant level.

In Experiment 6, we focused our analysis on free recall rather than recognition. Participants were randomly assigned to a crime scenario involving burglary, domestic violence, motor incident, or harassment. Female victims were involved in the motor incident and domestic violence scenarios, and male victims were involved in the burglary and harassment scenarios. Participants listened to the crime scenario and after a ten-minute break, they were asked to write the details of the crime, including the type of crime, date, time, location, responding officer, victim, and description of the crime that occurred. We explored whether male or female participants would recall more crime scenario details when the victims were female rather than male and examined whether female participants would recall more details across all crime scenarios than male participants. We analyzed the effect of the victim's gender on recall accuracy while controlling for the crime scenario condition and the effect of the participant's gender on recall accuracy controlling for victim's gender and found that neither the effect of the victim's gender or participant's gender were statistically significant.

The overall findings and data trends in this thesis show that earwitness identification is flawed and eyewitness identification strategies are not ideally transferable to auditory modalities. Recognition of stimuli presented during the encoding phase did not impact accuracy. Gender differences among speakers and alterations in stimuli content did not enhance identification; however, changes in duration significantly increased accuracy. This thesis examined theoretical paradigms and speaker identification strategies to contribute useful applications to reduce misidentifications; however, efforts to explore speaker identification and develop standardized lineup procedures should continue.

Encoding specificity principle

For Experiments 1 to 5, participants performed a recognition task that included previously heard stimuli integrated with new stimuli. The *encoding specificity principle* (Tulving & Thomson, 1973) suggests that participants should accurately recognize stimuli that were initially presented during the encoding phase. While the data trend in Experiment 1 shows that participants did recognize the same words and voices that they heard during encoding, there were high false alarms and the effects were not statistically significant. Our results for Experiments 1 do not support the findings on previous research involving old/new task encoding. In Experiment 2, the same stimuli presented in Experiment 1 was presented. Participants did recognize the original words presented during the encoding phase with very high accuracy but the false alarm rates were also high. The results found that participants were able to distinguish between old and new words to a significant level based on the effect of the speaker's gender but the small sample size does not adequately support the finding.

Participants' performance in Experiment 3 suggested a trend effect of discrimination between old and new voices and old and new phrases. We tested the recognition of voices and neutral content phrases presented during the encoding phase, and the hit rates for the speaking voice samples showed that participants could identify speakers to a significant level, but this result did not reflect d' prime sensitivity recalculation. When we analyzed the effect of content duration on accuracy, the result was statistically significant. Typically, neutral content is remembered less accurately than obscene material (Leander, Granhag, & Christianson, 2005) and neutral material could impair encoding for later recognition. Therefore, we changed both the content and presentation format in Experiment 4 to incorporate both written and auditory formats. Participants were presented with several statements that described crime scene details and completed a recognition test to determine how accurately they recognized original statements that were presented during the encoding phase compared to altered versions of those statements. Hit rates for the written format were higher than the auditory format but the effect was not statistically significant. Less cognitive resources are employed when participants recall details in a written format rather than a spoken format (Fueller et al., 2013) but our findings do not support the written superiority effect.

In Experiment 5, the same stimuli from Experiment 3 were presented with an additional rating for familiarity and confidence. The familiarity of the speakers and participants' confidence ratings did not significantly predict accuracy. Participants tend to assign markers to familiar voices during encoding and, when they recognize those markers, they identify that voice as familiar (Yarmey, 2012). The data trend showed that participants were somewhat to moderately familiar with the speakers and, therefore, encoding markers were not likely prevalent for most of the voice samples.

In Experiment 6, the participants were given a crime scenario to read and typed a crime report detailing the incident. The information was not presented more than once to support the encoding specificity principle. Arguably, details that are written down tend to include more correct information than those orally spoken (Sauerland et al., 2014). Therefore, it could be possible that details of the scenario were processed differently by the participants with some participants deeply processing the details but the levels of processing for each participant would be difficult to determine (Craik & Lockhart, 1972 as cited in Mandler & Johnson, 1977).

Voice Sample Duration

Previous research suggested that exposure to voice samples of 30 seconds or longer led to an increase in speaker identification accuracy (Kerstholt et al., 2004). Experiments 1 and 2 used voice samples of approximately 2 seconds long and voice samples in Experiments 3, 4, 5 and 6 were increased to approximately 30 seconds. Research has proven that shorter durations of only a few seconds could lead to speaker identification accuracy levels above chance (Bricker & Pruzansky, 1966, as cited in Yarmey, 2012). Manzanero and Barón (2017) tested participants' ability to recognize voices in a lineup after hearing several two-second voice samples. Participants recognized over 83% of the target voices; however, false alarms were over 50% in the target-absent lineup. In Experiments 1 and 2, participants heard short duration voice samples implemented speaker identification with short duration voice samples. In Experiment 1, the participants did not identify the speakers' voices to a significant level.

To improve encoding issues with short duration stimuli, we added a crime scenario of a physically violent robbery in Experiment 2. Previous research showed

that participants remembered content that included sexual or violent details with a higher accuracy level than neutral content (Pezdek & Prull, 1993). We anticipated that the provocative content would improve the encoding of the voices and increase accuracy levels. The data trends showed high hit rates for the male or female speaker portraying the robber in the scenario but false alarms were over 60% for male participants. The results found a significant main effect for speaker gender on speaker identification.

Typically, what is essential to any witness in any particular event is the amount of exposure they have to that sensory information. It is understood that although exposure as short as two seconds can still be recognized, exposure of a longer duration tends to increase identification accuracy and also reduce false alarms (Cook & Wilding, 1997b). In Experiments 3, 4, and 5, we increased the duration of the voice samples to ~30 seconds. According to Kerstholt et al. (2004) participants did not perform much better with a longer exposure duration of 70s (46%) compared to a shorter exposure duration of 30s (38%). Only after a lengthy retention interval did exposure duration make a difference in accuracy where performance was better for those exposed to the voice for a longer duration than a short duration. Yarmey (1995) found that identification accuracy was better in target-present and target-absent lineups when the participants heard the voice sample for 8 minutes compared to 30 seconds and showed that false alarm rates did not increase with the longer duration as reported in other studies. In Experiments 4 and 5, the results were not significant; however, in Experiment 3, we compare the short duration words from Experiment 1 with the long duration phrases in Experiment 3 and found a significant main effect of duration.

In addition to longer exposure times, additional presentations of the voice

sample greatly improved accuracy rates. At least three sessions of exposure to a voice sample greatly enhanced voice identification performance (Goldstein & Chance, 1985). Specifically, extensive exposure of voice samples showed better performance and accuracy rates than forewarning participants to prepare for an impending memory test (Deffenbacher et al., 1989). Previous research has shown that prior notification of a recognition test leads to better speaker identification accuracy than tests without forewarning. Though preparation is undoubtedly helpful, it is improbable that forewarning will occur in real-life settings.

Source confusion/misinformation

Schemas are constructs that organize our memories but they can create distortion within the memory construct by interfering the encoding and retrieval (Holt, 2019). After a memory is constructed, information presented afterward may cause confusion when recalling the accurate details of initial memory, thereby misattributing the correct source (Johnson, 1997). In Experiment 4, participants read or heard several crime scene statements and completed a recognition test to detect original and altered statements. Typically, when witnesses are exposed to an event and, later, presented with conflicting information, they are not able to correctly attribute the memory of the event to initial source; rather they misattribute the memory to the newly acquired information source (Holt, 2019). The participants were able to distinguish the original statements from those that were altered with high accuracy. The data trends conflict with source confusion research findings but our results were not statistically significant.

Familiarity

Strong speaker familiarity does not necessarily yield high identification accuracy (Read & Craik, 1995). In Experiment 5, we introduced a familiarity rating

for the stimuli we initially presented in Experiment 3. Participants heard ten voices spoken by males and females during an interview segment. Across all speakers, participants rated their familiarity to be somewhat familiar to moderately familiar. The data trend showed that familiarity with the speakers' voices did not predict accuracy for identification. Yarmey et al. (2001) analyzed familiarity ratings on a scale of unfamiliar, low, moderate, and high levels. Participants performed better in identifying voices that they rated as moderately to high familiarity compared to those rated as low to unfamiliar. Unfortunately, the results are mixed because higher accuracy of hit rates also tends to follow with higher false alarm rates (Kerstholt et al., 2004).

Familiar and unfamiliar accents also impact identification accuracy. The “other-accent” effect displays similarities with the “other-race” effect in eyewitness identification where witnesses tend to identify faces of the same race at a higher rate than those of another race (Wright & Sladden, 2003). The “other-accent” effect is established on the same principles that suggest people can identify people with the same or familiar accent than a different or unfamiliar accent. Stevenage et al. (2012) compared identification accuracy among English and Glaswegian participants exposed to English and Glaswegian accents. While Glaswegian are more likely to be exposed to English accents, the opposite was much less probable. In a target-present lineup, English participants identified English accents better than Glaswegian accents and vice versa. It is essential to note that, although we did not analyze the “other-accent” effect, efforts were made to ensure that, across all experiments in this thesis, participants were exposed to voice samples that exhibited neutral, South England accents. Still, it is likely that some participants may not have had a strong familiarity with South England accents but considering most participants were students at a UK

higher education institution, some familiarity with the South England accent would exist.

Confidence

Confidence rating measures are crucial in eyewitness identification research and a similar application of confidence can be found in earwitness identification. The impact of confidence ratings in court testimony is high and jurors tend to give credence to identifications reflecting higher confidence ratings (Howe et al., 2017). In Experiment 5, we analyzed participants confidence ratings as a predictor of accuracy. The data trends did not find relationship between confidence ratings and speaker identification accuracy. Similarly, Öhman et al. (2011), did not find that a correlational relationship between confidence and accuracy existed. However, Wixted, Read, and Lindsay (2016) found that participants with higher confidence ratings provided more accurate identifications over various retention intervals. This suggests that the findings for the confidence-accuracy relationship are mixed and caution should be taken when relying on confidence as evidentiary support.

Writing superiority effect

When witnesses come forward to report a crime, the reports can be completed in written format or orally. As law enforcement moves towards utilizing the digital space through the internet and social media, the traditional channels of crime reporting may shift. Young adolescents, also known as the “mobile youth culture,” use mobile devices more frequently than their older counterparts (Vanden Abeele, 2016). While mobile devices provide access to information and communication mediums, more options to shop, collaborate with colleagues, and display creativity are freely accessible. The foray into crime reporting via smartphone has become widely

available in forward-thinking countries like Australia (Paterson et al, 2018). Crime reports can be generated through type-written or oral formats.

Furthering this objective, Experiment 6 implemented the use of a mock crime report to determine how well participants can recall details of a crime. In Experiment 6, participants read a short excerpt that depicted a crime of burglary, domestic violence, harassment, or motor incident. They took a short break and then completed a crime report that asked them to recall the crime details. The participants typed their responses and the details were scored based on the correct idea units recalled. Males performed better on the recall task than females but the results were not statistically significant.

Sauerland and Sporer (2011) suggest that most witnesses will be interviewed orally by law enforcement, thus giving an oral account of a crime. While oral interviews may still occur, efforts to streamline paperwork and investigative procedures may require digital delivery of crime details rather than a formal interview. The writing superiority effect suggests that a written account rather than an oral account of an event, results in a more accurate recall of the event details (Sauerland et al., 2014). Researchers support this modality because it uses less cognitive resources to perform the writing task (Grabowski, 2005). Employing a written modality may become more common as the mobile youth culture advances technological innovation in communication, social interaction, and community safety.

8.2 Practical limitations

In the present thesis, there are some limitations to consider. Experiment 2 yielded a statistical significance for the speakers' gender, but a larger sample size would reflect increased power and effect size for this result to adequately support

previous research. Across all experiments, the smaller sizes led to non-significant results and future testing should account for small sizes to ensure statistical power of at least 0.8. Alternatively, multilevel modelling to partial out stimulus effects could be conducted with the current data to increase power analyses but that warrants a statistical sophistication to explore at a later time.

Changes in the Retention Interval

As we have typically seen in real-life situations, there is a likelihood of a delay between witnessing a criminal event and providing identification of the perpetrator. Experiments 1, 2, 3, 4, and Pilot studies 1 and 2 implemented the content recognition or speaker identification test after a short delay of hearing the voice samples or reading written content. Experiment 5 implemented a speaker identification test after a delay of 10 minutes from initially hearing the voice sample. Experiment 6 implemented a recall test after a delay of 10 minutes from initially hearing the voice sample. Although these delays were constructed within the experiments, we did not intentionally manipulate these delays to determine the effect of retention intervals on speaker identification and content accuracy.

Previous research has shown that longer delays of at least three weeks have had a negative effect on identification accuracy (Yarmey and Matthys, 1992). It is likely that several weeks, months, and, even years, would be a typical delay that witnesses face from the initial event to providing identification evidence. As we saw in the Hauptmann trial, three years passed before Charles Lindbergh gave testimony on recognizing the voice of the accused after only hearing the perpetrator speak two words. Such extensive passage of time may lead to more recognition inaccuracies and warrants further exploration.

Conducting a Lineup

Voice lineups are usually conducted as a target-present or target-absent format (Brewer, Weber, Wootton, & Lindsay, 2012). The voice samples in voice lineup can be presented in a serial or sequential order (Smith et al., 2020). A serial lineup introduces several voice samples in succession and then the witness identifies the perpetrator from the sequence of voice samples. By presenting all the voice samples before making an identification allows the witness to compare each of the samples to the lineup. The other option is to conduct a sequential lineup where each voice sample is presented to the witness and the witness gives a response after each sample until a positive identification is made. The sequential lineup reduces the obligation to make a selection because the witness could proceed through each voice until the final voice in the lineup and not select any of the voices. We did not conduct a lineup in our speaker identification experiments. Participants were presented with voice samples in sequential order but not in a lineup presentation. Further exploration of implementing a sequential lineup presentation may produce significant results that align with current research strategies for speaker identification.

Impact of Aural Characteristics

Auditory characteristics like tone, pitch, speaking rate, and amplitude have impacted how accurately people can recognize voices. Fluctuations in tone and pitch can alter the speaking voice from the initial exposure to later identification when those fluctuations have subsided. To reduce tonal and pitch changes, we focused on neutral South East English accents with no elements of emotionality and at a consistent volume across all experiments. Sensory memory focuses on specific auditory characteristics. By presenting the same voices spoken in the same neutral tones, we attempted to reduce or eliminate as much speaker variability as possible. Undetectable fluctuations could have occurred but they would have little effect on recognition

accuracy because research has suggested that difference in emotionality and neutral tones did not impact recognition accuracy (Saslove & Yarmey, 1985).

Age Differences

Age differences may impact speaker identification. Although very young children can recognize the voices of familiar speakers, adolescents and adults perform better when recognizing familiar and unfamiliar voices (Yarmey, 2012). Between the ages 21 to 40, witnesses are able to identify voices in lineup better than younger children and the elderly (Yarmey, 2012). We did not analyze the participants' ages but further consideration of age differences in the experiments could reveal new details about speaker identification.

8.3 Suggestions for Future Research

Earwitness misidentifications can lead to innocent individuals being selected in a lineup instead of the actual perpetrators. Speaker identification evidence is still held in high regard by law enforcement worldwide as an accurate measure of identification, much like the reliability of eyewitness identification (Hollien, 2012). However, unlike the extensive study of eyewitness identification, scrutiny of earwitness identification remains limited (Robson, 2018). Expert testimony on reliability is often dismissed and considered inadmissible because most judicial representatives presume that juries can use common sense to deduce suggestibility in eyewitness identification (Laub, Wylie, & Bomstein, 2013). In light of our present results, voice identification is malleable, and more stringent standards should be implemented when obtaining earwitness evidence.

A set of standards or guidelines for voice lineup procedures is essential to maintain objectivity and fairness. The U.S. does not currently have a comprehensive

guide for conducting voice lineups but several researchers have suggested guidelines based on empirical evidence that may lead to a standardized application in the future. Hollien (2012) recommends that the voice lineup should be comprised of 6-8 voices made up of foils and the perpetrator for a total of up to 25 samples. The samples should range from 1 -2 minutes and all voices should be presented to the witness before they decided to select or not select a voice. In lineup procedures, the perpetrator is not always in the lineup but, witness may erroneously identify an innocent person to just make an identification. Hollien (2012) suggests that the witness is told the perpetrator may or may not be in lineup as another alternative to eliminate obligatory identification of a foil. To reduce error and false alarms, it is suggested that law enforcement offers a “Don’t know” or “not present” option to witnesses as a plausible alternative so they do not feel a force obligation to make an identification of whomever may be present (Sanders, & Warnick, D., 1980).

A comprehensive guide detailing recommendations for eyewitness identification procedures (Wells et al., 2020) could be applied to earwitness identification until more aural-specific standards are developed. The recommendations build on Hollien’s (2012) suggestions and further recommend: (1) witnesses give a description of the suspect and details surrounding the crime, (2) law enforcement should have strong evidence to suspect that a suspect is guilty before including them in the lineup, (3) neither the person conducting the lineup nor the witness should know who the suspect is, (4) the prelineup instructions should indicate: (a) the administrator is blind to the lineup, (b) witness will say how confident they are in their selection immediately after, and (c) continue the investigation if the witness did not make an ID, (5) the whole procedure should be recorded, (6) do not present the same suspect to the same witness, (7) avoid showups and try to conduct a lineup.

In the UK, the Home Office (2003) developed standard policy guidelines for voice conducting voice parades and the aspects of the policy take into account the nuances that occur in speaker identification. Like Wells et al. (2020), they suggest a preparade interview to get a description of the suspect's voice in order to generate 20 foil samples. Once the foils are made, a linguist reviews the samples to make sure there are no distinguishable characteristics. The voice parade consists of 9, one-minute voice samples, including the suspect and foils. Similar to Hollien's (2012) suggestion, the witness should listen to all the samples before making a decision. They go further to reduce the retention interval by presenting a voice parade within 4-6 weeks of witnessing the event.

While some standard guidelines for voice lineups exists, they still fall short in several areas that we tested in our thesis. The lineup recommendations do not take into account the witness's familiarity with the suspect's voice, the duration of time that the witness heard the suspect, the suspect's gender, any content material that should be reported, and the modality of the prelineup interview (written or spoken). These factors have shown to impact accuracy and the current guidelines should consider sufficient strategies to incorporate them into future voice lineup recommendations.

New Technological Developments

The information that we gathered from our research has shown that memory for speaker identification is poor even when tested within controlled laboratory conditions. We did not find that participants' confidence supported speaker identification accuracy any more than recognition scores gathered without confidence ratings. Efforts to achieve more accurate identification have turned digital and law

enforcement agencies in the UK (The Smart Way to Report Crime, 2019), Thailand (Vatanasuk et al., 2015), and other countries have implemented new technological strategies. Recent digital developments have streamlined identification evidence. In the last decade, video lineups have created a straightforward system to conduct lineups in and outside of the police station (Memon, Havard, Clifford, Gabbert, & Watt, 2011). Although there appears to be no change in identification accuracy in video lineups and live lineups, the latter offers witnesses a convenient alternative and possibly, and a shorter retention interval duration.

Moving to a digital reporting system will likely lead to fewer errors in captured and recalled evidence and, successfully reduce wrongful convictions of innocent men and women. These strategies will reduce extra paperwork that will allow for a more streamlined process to accurately report crimes and quickly deliver those reports to officials through electronic means and, thus, unburdening flawed paper and telephone system services. As smartphone users continue to increase, the need to streamline eyewitness accounts and incident reports is imperative.

Researchers in Thailand (Vatanasuk et al., 2015) and Australia (Paterson et al., 2018) are seeing the need for more digital programs like mobile applications that will capture eyewitness evidence. A digital tool may prove to be invaluable to government agencies, law enforcement, and legal professional as they continue to tackle crime and prosecute the offenders. The current burden on the legal system has not offered any resolutions to reduce identification inaccuracies. Organizations like the Innocence Project (The Innocence Project, 2014), while noble in effort, are tirelessly working on the backend to exonerate individuals who have been wrongfully incarcerated. As hopeful as their progress is, it is not enough. I am hoping, that in

some small measure, my mobile application will contribute to the evolving security needs of society and safeguard justice.

Appendices

Appendix A – Experiment 2 Crime Scenario

Male Attendant Scenario

A woman goes into a local shop armed with a handgun. She kills time by looking at the drink selection in the fridge until the last customer leaves the shop. She covers her mouth with a bandana and approaches the attendant. She holds him at gunpoint and demands that he empties the till into a bag that she brought. The attendant grabs a nightstick from under the counter and swings at her chasing her off.

Male Robber Scenario

A man goes into a local shop armed with a handgun. He kills time by looking at the drink selection in the fridge until the last customer leaves the shop. He covers his mouth with a bandana and approaches the attendant. He holds her at gunpoint and demands that she empties the till into a bag that he brought. The attendant grabs a nightstick from under the counter and swings at him, chasing him off.

Appendix B – Experiment 4 Old/Altered Statements

[illegible]

--	--

**Note: Audio statements did not include Statements 19 and 20*

Appendix C - Experiment 5 Familiarity and Confidence testing

Welcome to the study!

The Effect of Cued Memory Recall Strategies on Word Identification and Voice Recognition

This survey is for experimental purposes only in conjunction with City, University of London, Department of Psychology.

How are you accessing this survey?

- ☐ I'm using a desktop computer (1)
 - ☐ I'm using a mobile device/tablet (2)
-

WELCOME TO PART 1!

YOU WILL HEAR 10 AUDIO CLIPS.
PLEASE LISTEN CAREFULLY.
NO FEEDBACK IS REQUIRED AT THIS TIME.

IF YOU HAVE EARPHONES OR HEADPHONES,
PLEASE PLUG THEM IN NOW.

ONCE YOU CLICK ON THE ARROW TO CONTINUE, THE RECORDINGS
WILL BEGIN.

Please rate the familiarity of the voice

- ☐ Extremely familiar (5)
 - ☐ Very familiar (4)
 - ☐ Moderately familiar (3)
 - ☐ Slightly familiar (2)
 - ☐ Not familiar at all (1)
-

YOU HAVE COMPLETED PART 1.

PLEASE TAKE A 10 MINUTE BREAK
BEFORE CONTINUING TO PART 2.

IF YOU TAKE LONGER THAN 10 MINUTES OR DO NOT RETURN TO
COMPLETE THE SURVEY, THEN YOUR RESULTS WILL BE VOIDED AND
NO PAYMENT WILL BE PROVIDED

WELCOME TO PART 2!

YOU WILL HEAR 20 AUDIO CLIPS.

YOU WILL BE ASKED TO DETERMINE
IF THE VOICE IS AN "OLD" VOICE
YOU PREVIOUSLY HEARD IN PART 1
OR A "NEW" VOICE.

PLEASE ANSWER EACH QUESTION.

NO PAYMENT WILL BE PROVIDED FOR AN INCOMPLETE SURVEY.

IF YOU HAVE EARPHONES OR HEADPHONES,
PLEASE PLUG THEM IN NOW.

ONCE YOU CLICK ON THE ARROW TO CONTINUE, THE RECORDINGS
WILL BEGIN.

IS THIS AN OLD VOICE FROM PART 1 OR A NEW VOICE?

☐ OLD (0)

☐ NEW (1)

BASED ON YOUR ANSWER IN THE PREVIOUS QUESTION, PLEASE RATE HOW CONFIDENT YOU ARE WITH YOUR ANSWER.

	1 = Not very confident (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 = Very confident (7)
Confidence rating (1)	☐	☐	☐	☐	☐	☐	☐

Appendix D – Experiment 6 - Provide the Proof Mobile Application Test

WELCOME TO PROVIDE THE PROOF

WELCOME TO PART 1!

EACH COLOUR REPRESENTS AN AUDIO RECORDING.

PLEASE SELECT **ANY COLOUR** FROM THE OPTIONS BELOW.

ONCE YOU CLICK ON THE ARROW TO CONTINUE, YOU WILL HEAR AN AUDIO RECORDING.

PLEASE LISTEN CAREFULLY.

- ☐ Red (1) – *corresponded to Domestic Violence*
- ☐ Blue (2) – *corresponded to Burglary*
- ☐ Green (3) – *corresponded to Motor Incident*
- ☐ Yellow (4) – *corresponded to Harassment*

Please listen to the entire message before selecting the arrow to continue

NO FEEDBACK IS REQUIRED

YOU HAVE COMPLETED PART 1.

PLEASE TAKE A 10 MINUTE BREAK
BEFORE CONTINUING TO PART 2.

IF YOU TAKE LONGER THAN 10 MINUTES OR DO NOT RETURN TO
COMPLETE THE SURVEY, THEN YOUR RESULTS WILL BE VOIDED AND
NO PAYMENT WILL BE PROVIDED

WELCOME TO PART 2!

YOU WILL COMPLETE A "CRIME REPORT" BASED ON THE AUDIO
RECORDING YOU PREVIOUSLY HEARD.

PLEASE ANSWER EACH QUESTION.

ANY QUESTION LEFT BLANK/INCOMPLETE WILL BE CONSIDERED AN INCOMPLETE REPORT.

NO PAYMENT WILL BE PROVIDED FOR AN INCOMPLETE REPORT.

PLEASE SELECT THE ARROW TO CONTINUE

Based on the audio recording you heard previously,
please select the TYPE OF CRIME that occurred:

- ☐ BURGLARY (1)
- ☐ DOMESTIC VIOLENCE (2)
- ☐ HARASSMENT (3)
- ☐ MOTOR INCIDENT (4)

What was the GENDER of the responding police officer(s)?

- ☐ Male (1)
- ☐ Female (2)

What was the GENDER of the victim(s)?

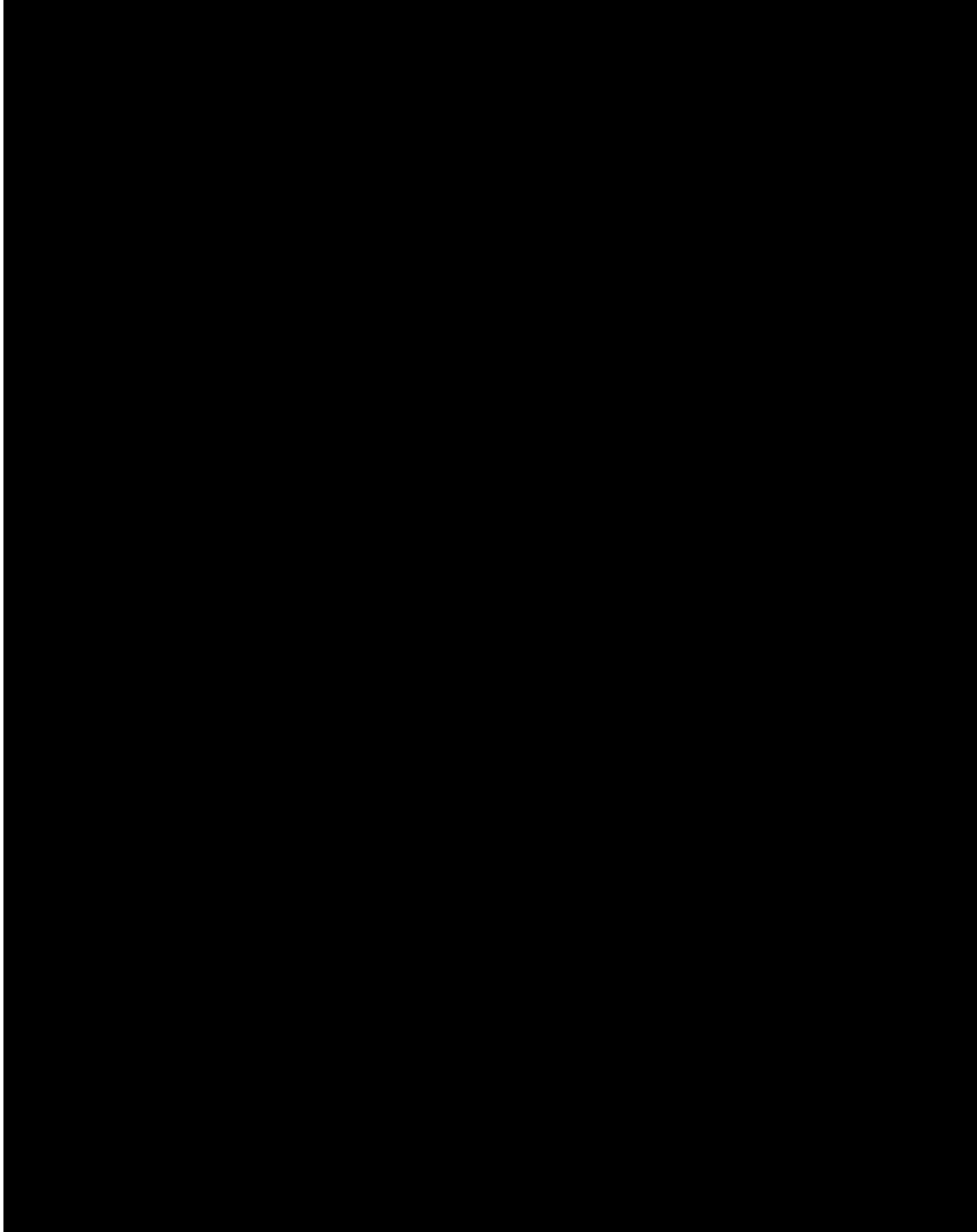
- ☐ Male (1)
- ☐ Female (2)

WHEN did the incident occur (i.e. date/time)?

WHERE did the incident occur (i.e. home)?

Please give a description of the crime committed and add any other factors that are relevant to the report.

Appendix E – Experiment 6 - Guide to scoring idea units (Mandler & Johnson, 1977) and Crime Scenarios (separated into idea units)



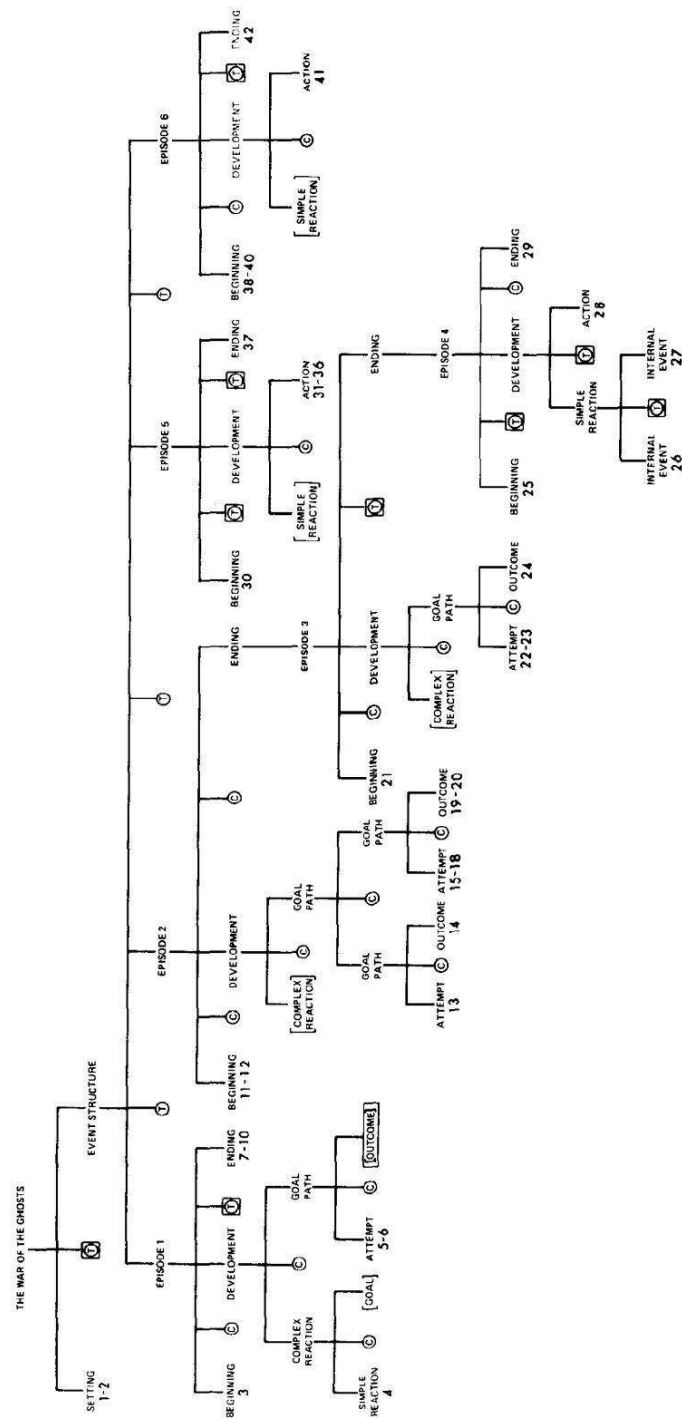


Fig. 5. A simplified representation of the War of the Ghosts story. Only basic nodes are shown. Violations are enclosed in boxes.

Burglary (15 idea units)

1. Date recall - Friday, April 9th, 2016

2. Time recall - at 9:45 am
3. Name of Officer, where and crime - I, Officer Janice Ross was dispatched to 21 Powell Street
4. to investigate a burglary.
5. Who - I met with Frank Gaines
6. the homeowner who had reported the burglary.
7. Gaines told me he lives alone.
8. He was out of town on business when the burglary happened.
9. He had left on Monday, April 5th at approximately 6:15 pm
10. and returned on Friday morning at approximately 8:45 am
11. and he used his car for the trip
12. so there was no car in his carport when he was gone.
13. When he returned from his trip
14. he saw a broken window over the kitchen table.
15. The following items are missing from his home office - a dell computer and a printer.

Domestic Violence (19 idea units)

1. Time recall - At 8:15 pm
2. Date recall - on January 4, 2017
3. Name of Officer, where, incident - I, Officer John Brown was dispatched
4. to a domestic disturbance at 30 Crown Place.
5. I knocked on the front door
6. and called out police officer.
7. I heard a woman's voice yell "I hate you, I hate you!"
8. I heard a man's voice yell "Shut up!"
9. No one answered the door.
10. I tried the door knob
11. and I entered the living room.
12. Who - A woman, Jane Brown was sitting on the sofa.
13. There was a red mark on her right cheek.

14. Her lips were trembling,
15. her face was wet
16. and her eye makeup was smeared.
17. Who - A man, Tim Brown was standing over her.
18. His fists were clenched.
19. He said he hit her and would do it again.

Harassment (15 idea units)

1. Time - At 6:30 pm
2. Date - on May 7, 2017
3. Officer, crime and where - I, Officer Sydney Taylor was dispatched
4. to a harassment call at 10 Green Lane.
5. Who - I saw two men arguing.
6. One man, Brad Johnson yelled "go back to where you came from."
7. Who - The other man Yusef Zand replied "shut up."
8. I approached both men
9. and separated them.
10. I told them both to calm down
11. and I asked them to explain what happened.
12. Yusef said that Brad pushed him when he walked past him.
13. Yusef then stopped
14. and started yelling
15. that he was going to call the police.

Motor Incident (14 idea units)

1. Time - At 5:42 pm
2. Date - on February 5, 2018
3. Officer - I, Officer Larry Smith was dispatched
4. to a motor incident.
5. Where, who and crime - The owner Ashley Daynia was towards her residence

6. while returning from her office
7. when a motor bike rider hit her at a blind corner
8. leaving her with minor injuries.
9. The car boot and corner window were badly damaged.
10. The motor bike rider escaped from the scene immediately.
11. Ashley managed to take a note of the motor bike plates number which is CGH 493.
12. I noted that Ashley had an injured knee
13. that occurred on impact.
14. No other injuries were detected.

Appendix F – Prototype Mobile Application

Prototype Mobile Application (Android version)

The login screen features a blue header with the text "Login". Below the header is a logo that reads "Provide the PROOF" in a stylized font. There are two input fields: "Email Address" and "Password". Below these fields are three buttons: a blue "LOGIN" button, a grey "REGISTER" button, and a grey "HELP" button with a question mark icon.

The home screen has a blue header with the text "Home". Below the header is a section titled "Start" with a welcome message: "Welcome to Provide the Proof, the easy way to report to your local police department. To start reporting click on **Report**. Find all your submitted reports in **My Archive**." Below this message are two buttons: a blue "REPORT" button and a grey "MY ARCHIVE" button.

The report screen has a blue header with a "BACK" button and the text "Report". Below the header is a text box with instructions: "Please complete all the sections of form below. Select the crimes you are reporting, confirm the location, add evidence and submit a description of the event." Below this is a section titled "Crime Selection" with a list of crimes and checkboxes: Burglary, Domestic Violence, Fraud, Harassment, Hate Crime, Motor Incident, Mugging, Sexual Offense, and Terrorism.

This part of the report screen continues from the previous one. It features a "LOGOUT" button at the top. Below it are four input fields: "Town", "County", "Postcode", and "Country". Below these fields is a section titled "Evidence Submission" with an "UPLOAD IMAGES" button. Below that is a section titled "Description" with a text input field labeled "Enter your description here...". At the bottom is a blue "SUBMIT" button.

Appendix G – Research Committee Ethics Application



Psychology Department Standard Ethics Application Form: Staff, PhD Students, MRes Students

This form should be completed in full. Academic staff should email it to psychology.ethics@city.ac.uk. Students and research assistants should email it to their supervisor who should approve it before submitting it to psychology.ethics@city.ac.uk. Please ensure you include the accompanying documentation listed in question 19.

Does your research involve any of the following? <i>For each item, please place a 'x' in the appropriate column</i>	Yes	No
Persons under the age of 18 <i>(If yes, please refer to the Working with Children guidelines and include a copy of your DBS)</i>		X
Vulnerable adults (e.g. with psychological difficulties) <i>(If yes, please include a copy of your DBS where applicable)</i>		X
Use of deception <i>(If yes, please refer to the Use of Deception guidelines)</i>		X
Questions about topics that are potentially very sensitive <i>(Such as participants' sexual behaviour, their legal or political behaviour, their experience of violence)</i>		X
Potential for 'labelling' by the researcher or participant (e.g. 'I am stupid')		X
Potential for psychological stress, anxiety, humiliation or pain		X
Questions about illegal activities		X
Invasive interventions that would not normally be encountered in everyday life (e.g. vigorous exercise, administration of drugs)		X
Potential for adverse impact on employment or social standing		X
The collection of human tissue, blood or other biological samples		X
Access to potentially sensitive data via a third party (e.g. employee data)		X
Access to personal records or confidential information		X
Anything else that means it has more than a minimal risk of physical or psychological harm, discomfort or stress to participants.		X

If you answered 'no' to all the above questions your application may be eligible for light touch review. We aim to send you a response within 7 days of submission. However, review may take longer in some instances, and you may also be asked to revise and resubmit your application. Thus you should ensure you allow for sufficient time when scheduling your research.

If you answered 'yes' to any of the questions, your application is NOT eligible for light touch review and will need to be reviewed at the next Psychology Department Research Ethics Committee meeting. These take place on the first Wednesday of every month (with the exception of January and August). Your application should be submitted at least 2 weeks in advance of the meeting you would like it considered at. We aim to send you a response within 7 days. Note that you may be asked to revise and resubmit your application so should ensure you allow for sufficient time when scheduling your research. If the research is considered very high risk, or the committee does not feel it has the expertise to review it, we may ask you to submit your application to the Senate Research Ethics Committee.

If you are unsure about any of above, please contact the Chair of the Psychology Department Ethics Committee, [REDACTED]

Is this project supported by external funding?	Yes	No
		X
If you answered yes, please provide the name of the funding body and the amount awarded.		

Which of the following describes the main applicant? <i>Please place a 'x' in the appropriate space</i>	
Undergraduate student	
Taught postgraduate student	
Professional doctorate student	
Research student	X
Staff (applying for own research)	
Staff (applying for research conducted as part of a lab class)	

1. Name of applicant(s).
Tiffany Lauren Elmore
2. Email(s).

3. Project title.
The Effect of Cued Memory Recall Strategies on Word Identification and Voice Recognition
4. Provide a lay summary of the background and aims of the research. (No more than 400 words.)
In the US, more than 70% people were wrongfully convicted largely due to inaccurate eyewitness identification. At present, both eyewitness and earwitness testimony is permitted in court trials. The questions remain as to how much reliance can be placed on the accuracy of this information and whether it should be regarded as a reliable evidence in police lineups and subsequent court testimony. The U.S. Supreme Court evaluated several factors that were critical to determine whether misidentification violated a defendant's right to due process under the Fourteenth Amendment of the U.S. Constitution based on criteria established in Neil v. Biggers (1972). The Court provided five factors that that determined the admissibility and reliability of eyewitness and, presumably, earwitness identification when confrontation procedures were deemed suggestive (Deffenbacher et al., 2001). The criteria are: [a] The opportunity of the witness to view the criminal, at the time of the crime, [b] the witness' degree of attention, [c] the accuracy of the witness' prior description of the criminal, [d] the level of certainty demonstrated by the witness at the time of confrontation, [e] the length of time between the crime and the confrontation (p. 199). Previous studies have examined whether extended exposure to an auditory stimulus increases voice identification accuracy. It was argued that extended exposure of a 162-word stimulus distributed over a set length of time showed a more accurate hit rate when identifying the speaker among a nine-voice lineup than when the same stimulus was presented once and retention was tested weeks later (Goldstein & Chance, 1985). The results varied with the length of the stimulus and also failed to reflect validity in real-life situations where only one utterance may be heard by a witness and extensive delays make recall of minimal utterances difficult.

The aim of the present study is to assess the validity of the following criteria as it applies to voice recognition: the opportunity of the witness to hear the speaker and the witness' degree of attention.

The study will examine:

- (1) How accurate are people in identifying voices?
- (2) Does accuracy vary with the gender of the identifier?
- (3) Does accurate identification of a voice lead to accurate identification of the word(s) spoken by the identified voice?
- (4) How does voice and word identification impact earwitness identification in real-life situations?

5. Provide a summary of the design and methodology.

Design

The overall study employs a 2 x 2 x 2 mixed design. The between-subjects variable is participant gender, with two levels (male or female). The within-subjects variable is voice gender, with two levels (female or male) and type of test, with two levels (voice recognition or word identification).

Participants

One hundred and forty undergraduate students from City University. All are fluent English speakers with no history of memory or hearing disorders. All subjects will be screened for normal hearing prior to the start of the study. Any subject exhibiting a hearing loss will be not be selected to continue participation in the study.

Apparatus and Materials

The stimuli are 269 monosyllabic words from the Modified Rhyme Test (MRT) and an additional 131 monosyllabic words from Egan's Articulation Testing Methods study. The list is constructed from a database of 400 words spoken by 20 speakers (10 males and 10 females). All words will be recorded on a digital recording device and uploaded into E-Prime computer software.

Procedure

Participants will be tested individually in three phases. In the first phase of the study, participants will be instructed to use headphones to listen to 40 pre-recorded words presented on computer software. A total of 20 words were recorded in various male voices and 20 words were recorded in various female voices. Some voices are presented more than once in the lineup. Each monosyllabic word ranges from 1s to 2s in length with up to 5s of response time provided between the presentation of each word. After listening to all 40 voices, participants will engage in a filled interval task for 2 minutes before continuing with the next phase of the study.

For the filled interval, participants will view and identify 3 ambiguous visual figures: (1) the Necker cube, the duck-rabbit illusion, and My Wife and My Mother-in-law illusion. Following the filled interval task, participants will proceed to the Phase 2.

In Phase 2, all participants are presented with 80 words (40 read by two new male and female voices and 40 read by an original voice from Phase 1). Participants will determine whether the voice was from the original list of voices heard in Phase 1 or a new voice.

Participants will be seated in front of a desktop computer. After hearing each pre-recorded word, participants will respond by using a mouse to select a button labelled “new” if the voice was judged new or a button labelled “old” if the voice was judged old. Participants will be instructed to respond as quickly and as accurately as possible and given a maximum of 5s to respond in each trial. If no response is made, that trial will not be recorded. Participants will be advised that they are not being tested for word identity.

After the listening task is completed, participants will engage in a two minute filled interval and identify selected visual illusions. Following the task, participants will continue with Phase 3 of the listening task.

In Phase 3, all participants will be presented with 80 words (40 read by two new male and female voices and 40 read by an original voice from Phase 1). Participants will determine whether the word was from the original list of words heard in Phase 1 or a new word.

Participants will be seated in front of a desktop computer. After hearing each pre-recorded word, participants will respond by using a mouse to select a button labelled new if the word was judged as new or a button labelled old if the word was judged as old. Participants will be instructed to respond as quickly and as accurately as possible and given a maximum of 5s to respond in each trial. If no response is made, that trial will not be recorded. Participants will be advised that they are not being tested for voice identity.

6. Provide details of all the methods of data collection you will employ (e.g., questionnaires, reaction times, skin conductance, audio-recorded interviews).

Stimulus presentation and data collection will be through E-Prime software installed on a desktop computer at a computer lab in the Rhind Building of City University London. Participants will be assigned a numerical code and their names will be stored separately from the data.

7. Is there any possibility of a participant disclosing any issues of concern during the course of the research? (e.g. emotional, psychological, health or educational.) Is there any possibility of the researcher identifying such issues? If so, please describe the procedures that are in place for the appropriate referral of the participant.

The study does not require the participants to reveal information of a sensitive nature and the procedure will not cause any psychological distress. However, if there are any behavioural indicators of distress from a participant, the experiment will be postponed. If deemed appropriate, in house support services will be offered to the participant and further evaluation

will take place to determine their suitability for continued participation in the research. If appropriate support cannot be offered by the service, my supervisor or me, the contact number for Student Support Services and other relevant bodies will be provided.

8. Details of participants (e.g. age, gender, exclusion/inclusion criteria). Please justify any exclusion criteria.

Participants will be selected to participate in the research study provided they meet the following criteria:

- (1) An undergraduate student at City University London, including male and female students of all ages
- (2) A fluent English speaker
- (3) Exhibit normal hearing
- (4) Not presently suffering from any memory related illnesses
- (5) Provide written consent to participate

Prior to start of the study, all participants will complete a brief hearing examination. All participants must meet minimum hearing requirements to proceed with the study. Participants with substantial hearing loss will be excluded from the study as minimal hearing standards are required to hear to auditory stimuli presented for evaluation.

9. How will participants be selected and recruited? Who will select and recruit participants?

Study recruitment flyers will be posted in the Rhind Building of City University London to recruit undergraduate students. Participants who respond to the flyer by contacting the researcher and expressing interest to participate will be selected to participate in the preliminary hearing evaluation. Participants who meet the minimum hearing standards will be selected to continue with the study.

10. Will participants receive any incentives for taking part? (Please provide details of these and justify their type and amount.)

Participants will not receive any incentives for taking part in the study.

11. Will informed consent be obtained from all participants? If not, please provide a justification. (Note that a copy of your consent form should be included with your application, see question 19.)

Informed consent will be obtained from all participants prior to the preliminary hearing evaluation.

12. How will you brief and debrief participants? (Note that copies of your information sheet and debrief should be included with your application, see question 19.)

Participants will be provided with an information sheet prior to the study and a debriefing sheet after their participation has concluded (see attached).		
13. Location of data collection. (Please describe exactly where data collection will take place.)		
All data collection will take place in the Rhind Building located at City University London, Northampton Square, London, EC1V 0HB UK		
13a. Is any part of your research taking place outside England/Wales?		
No	☒	
Yes	☐	If 'yes', please describe how you have identified and complied with all local requirements concerning ethical approval and research governance.
13b. Is any part of your research taking place <u>outside</u> the University buildings?		
No	☒	
Yes	☐	If 'yes', please submit a risk assessment with your application or explain how you have addressed risks.
13c. Is any part of your research taking place <u>within</u> the University buildings?		
No	☐	
Yes	☒	If 'yes', please ensure you have familiarised yourself with relevant risk assessments available on Moodle.
14. What potential risks to the participants do you foresee, and how do you propose to deal with these risks? These should include both ethical and health and safety risks.		
The study does not involve any physical and social risks to the participants. The study does not require the participants to reveal information of a sensitive nature and the procedure will not cause any psychological distress. However, if there are any behavioural indicators of distress from any participant, the intervention will be postponed. If deemed appropriate, in house support services will be offered to them and further evaluation will take place to determine their suitability for continued participation in the research.		
15. What potential risks to the researchers do you foresee, and how do you propose to deal with these risks? These should include both ethical and health and safety risks.		
The study does not involve any physical and social risks to the researchers. All information provide to the researcher will be kept confidential. Only members of the research team will have access to it. All data collection, storage and processing will comply with the principles of the Data Protection Act 1998 and the EU Directive 95/46 on Data Protection. Under no circumstances will identifiable responses be provided to any other third party. Information emanating from the evaluation will only be made public in a completely unattributable format or at the aggregate level in order to ensure that no participant will be identified. However, should a participant disclose information that may result in a participant or anyone else being put at risk of harm, the researcher will take steps to inform the appropriate authorities. If this situation arises, the research team will discuss all possible options for ourselves and the participant before deciding whether or not to take any action.		

16. What methods will you use to ensure participants' confidentiality and anonymity? (Please note that consent forms should always be kept in a separate folder to data and should NOT include participant numbers.)		
<i>Please place an 'X' in all appropriate spaces</i>		
Complete anonymity of participants (i.e. researchers will not meet, or know the identity of participants, as participants are a part of a random sample and are required to return responses with no form of personal identification.)		
Anonymised sample or data (i.e. an <i>irreversible</i> process whereby identifiers are removed from data and replaced by a code, with no record retained of how the code relates to the identifiers. It is then impossible to identify the individual to whom the sample of information relates.)		
De-identified samples or data (i.e. a <i>reversible</i> process whereby identifiers are replaced by a code, to which the researcher retains the key, in a secure location.)	X	
Participants being referred to by pseudonym in any publication arising from the research		
Any other method of protecting the privacy of participants (e.g. use of direct quotes with specific permission only; use of real name with specific, written permission only.) <i>Please provide further details below.</i>		
17. Which of the following methods of data storage will you employ?		
<i>Please place an 'X' in all appropriate spaces</i>		
Data will be kept in a locked filing cabinet		X
Data and identifiers will be kept in separate, locked filing cabinets		X
Access to computer files will be available by password only		X
Hard data storage at City University London		X
Hard data storage at another site. <i>Please provide further details below.</i>		
18. Who will have access to the data?		
<i>Please place an 'X' in the appropriate space</i>		
Only researchers named in this application form		
People other than those named in this application form. <i>Please provide further details below of who will have access and for what purpose.</i>		X
Tiffany Lauren Elmore (applicant) Prof. Martin Conway (supervisor) – access is granted should any issues arise prior to, during or after the completion of the study		
19. Attachments checklist. *Please ensure you have referred to the Psychology Department templates when producing these items. These can be found in the Research Ethics page on Moodle.		
<i>Please place an 'X' in all appropriate spaces</i>		
	Attached	Not applicable
*Text for study advertisement	X	
*Participant information sheet	X	
*Participant consent form	X	
Questionnaires to be employed		X
Debrief	X	
Copy of DBS		X
Risk assessment		X
Others (please specify, e.g. topic guide for interview, confirmation letter from external organisation)		X

20. Information for insurance purposes.

(a) Please provide a brief abstract describing the project

In the US, more than 70% people were wrongfully convicted largely due to inaccurate eyewitness identification. At present, both eyewitness and earwitness testimony is permitted in court trials. The U.S. Supreme Court evaluated several factors that were critical to determine whether misidentification violated a defendant's right to due process under the Fourteenth Amendment of the U.S. Constitution based on criteria established in Neil v. Biggers (1972). The Court provided five factors that that determined the admissibility and reliability of eyewitness and, presumably, earwitness identification when confrontation procedures were deemed suggestive. The aim of the present study is assess the validity of the following criteria as it applies to voice recognition: the opportunity of the witness to hear the speaker and the witness' degree of attention. The study tested the recall of 400 monosyllabic words presented in both male and female voices to randomly selected male and female participants. The analysis evaluated the participants' memory recall accuracy of both voices and words and noted differences between the gender of the participants and the gender of the voices. A hit was defined as responding "old" to an original voice or word and a false alarm as responding "old" to a new voice or word. The evaluation of hits and false alarms determined the accuracy of memory recall for words and voices after a short delay. The results will suggest that cued memory recall strategies will impact word identification and voice recognition accuracy.

Please place an 'X' in all appropriate spaces

(b) Does the research involve any of the following:	Yes	No
Children under the age of 5 years?		X
Clinical trials / intervention testing?		X
Over 500 participants?		X
(c) Are you specifically recruiting pregnant women?		X

(d) Is any part of the research taking place outside of the UK?		X
If you have answered 'no' to all the above questions, please go to section 21. If you have answered 'yes' to any of the above questions you will need to check that the university's insurance will cover your research. You should do this by submitting this application to [REDACTED], <u>before</u> applying for ethics approval. Please initial below to confirm that you have done this. I have received confirmation that this research will be covered by the university's insurance. Name Date.....		

21. Information for reporting purposes.		
<i>Please place an 'X' in all appropriate spaces</i>		
(a) Does the research involve any of the following:	Yes	No
Persons under the age of 18 years?		X
Vulnerable adults?		X
Participant recruitment outside England and Wales?		X
(b) Has the research received external funding?		X

22. Declarations by applicant(s)		
<i>Please confirm each of the statements below by placing an 'X' in the appropriate space</i>		
I certify that to the best of my knowledge the information given above, together with accompanying information, is complete and correct.		X
I accept the responsibility for the conduct of the procedures set out in the attached application.		X
I have attempted to identify all risks related to the research that may arise in conducting the project.		X
I understand that no research work involving human participants or data can commence until ethical approval has been given.		X
	Signature (Please type name)	Date
First applicant	[REDACTED]	4/12/2015
Supervisor (For students and research assistants only. Please ensure the <u>supervisor</u> submits the form.)	[REDACTED]	4/12/2015

Appendix H – Light Touch Ethics Approval



CITY UNIVERSITY
LONDON

Psychology Research Ethics Committee
School of Arts and Social Sciences
City University London
London EC1R 0JD

26th January 2016

Dear Tiffany Elmore

Reference: PSYETH (R/L) 15/16 145

Project title: The Effect of Cued Memory Recall Strategies on Word Identification and Voice Recognition

I am writing to confirm that the research proposal detailed above has been granted approval by the City University London Psychology Department Research Ethics Committee.

Period of approval

Approval is valid for a period of three years from the date of this letter. If data collection runs beyond this period you will need to apply for an extension using the Amendments Form.

Project amendments

You will also need to submit an Amendments Form if you want to make any of the following changes to your research:

- (a) Recruit a new category of participants
- (b) Change, or add to, the research method employed
- (c) Collect additional types of data
- (d) Change the researchers involved in the project

Adverse events

You will need to submit an Adverse Events Form, copied to the Secretary of the Senate Research Ethics Committee [REDACTED], in the event of any of the following:

- (a) Adverse events
- (b) Breaches of confidentiality
- (c) Safeguarding issues relating to children and vulnerable adults
- (d) Incidents that affect the personal safety of a participant or researcher

Issues (a) and (b) should be reported as soon as possible and no later than 5 days after the event. Issues (c) and (d) should be reported immediately. Where appropriate the researcher should also report adverse events to other relevant institutions such as the police or social services.

Should you have any further queries then please do not hesitate to get in touch.

Kind regards

[REDACTED]

[REDACTED]

Appendix I – Light Touch Ethics Amendment

Psychology Department Research Ethics Committee

Project Amendments/Modifications Request for Extension

For use in the case of all research previously approved by City University London Psychology Department Research Ethics Committee.

Was the original application reviewed by light touch?

If yes, please send this form to the individual who reviewed the original application. Once they have approved the amendment and signed the form, it should be emailed to psychology.ethics@city.ac.uk

Was the original application reviewed at a full committee meeting?

If yes, please email this form to psychology.ethics@city.ac.uk. It will be reviewed by the committee chair.

Note that you only have to respond to the sections relevant to you.

Details of Principal Investigator and Study

Name	Tiffany Elmore
Email	[REDACTED]
Title of study	The Effect of Cued Memory Recall Strategies on Word Identification and Voice Recognition
REC reference number	PSYETH (R/L) 15/16 145

Study Duration

Start Date	26 January 2016
End Date	26 January 2019

Project Amendments / Modifications

Type of modification/s (tick as appropriate)

Research procedure/protocol (including research instruments)	X
Participation group	X
Information Sheet/s	
Consent form/s	
Other recruitment documents	
Sponsorship/collaborations	
Principal investigator/supervisor	
Extension to approval needed (extensions are given for one year)	
Other	

Details of modification (give details of each of the amendments requested, state where the changes have been made and attach all amended and new documentation)

Research Procedure: Part 1: Participants hear 20 audio clips, presented randomly for each speaker (10 male speakers, 10 female speakers – no

feedback provided), the clips are followed by a visual filled task of looking at a picture and counting the number of people hidden in the background.

Part 2: Participants hear 40 audio clips (20 altered audio clips, 10 males, 10 females and 20 clips from Part 1, 10 males, 10 females). They are asked which audio clip is correct.

Participation group: Participants will be selected to participate in the research study provided they meet the following criteria:

- (6) A fluent English speaker
- (7) Exhibit normal hearing
- (8) Not presently suffering from any memory related illnesses
- (9) Provide written consent to participate

Participants will receive course credit or monetary compensation for taking part in the study.

Participant recruitment: Study recruitment flyers will be posted in the Rhind Building of City University London to recruit undergraduate students. Public participants will be recruited through City SONA online recruitment database.

Justify why the amendment/extension is needed (including the period of extension being requested)

Changes were made to the research procedure and participant incentive.

Period of extension requested

Other information (provide any other information which you believe should be taken into account during ethical review of the proposed changes)

Change in the study team

Staff member

<i>Title, Name & Staff Number</i>	<i>Post</i>	<i>Dept & School</i>	<i>Phone</i>	<i>Email</i>	<i>Date and type of CRB disclosure*</i>

Student

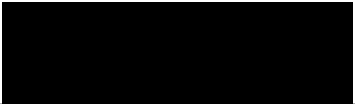
<i>Name & Student Number</i>	<i>Course / Year</i>	<i>Dept & School</i>	<i>Date and type of CRB disclosure*</i>

External co-investigator/s

<i>Title & Name</i>	<i>Post</i>	<i>Institution</i>	<i>Phone</i>	<i>Email</i>	<i>Date and type of CRB disclosure*</i>

Declaration (to be signed by the Principal Investigator)

- I certify that to the best of my knowledge the information given above, together with any accompanying information, is complete and correct and I take full responsibility for it.

Principal Investigator(s) (student and supervisor if student project)		
Date	14/3/2017	

Reviewer signature

To be completed upon FINAL approval of the amendment.

	Signature (Please type name)	Date
Reviewer		

Appendix J – Light Touch Ethics Amendment

Psychology Department Research Ethics Committee

Project Amendments/Modifications

Request for Extension

For use in the case of all research previously approved by City University London Psychology Department Research Ethics Committee.

Was the original application reviewed by light touch?

If yes, please send this form to the individual who reviewed the original application. Once they have approved the amendment and signed the form, it should be emailed to psychology.ethics@city.ac.uk

Was the original application reviewed at a full committee meeting?

If yes, please email this form to psychology.ethics@city.ac.uk. It will be reviewed by the committee chair.

Note that you only have to respond to the sections relevant to you.

Details of Principal Investigator and Study

Name	Tiffany Elmore
Email	[REDACTED]
Title of study	The Effect of Cued Memory Recall Strategies on Word Identification and Voice Recognition
REC reference number	PSYETH (R/L) 15/16 145

Study Duration

Start Date	26 January 2016
End Date	26 January 2019

Project Amendments / Modifications

Type of modification/s (tick as appropriate)

Research procedure/protocol (including research instruments)	X
Participation group	X
Information Sheet/s	X
Consent form/s	X
Other recruitment documents	
Sponsorship/collaborations	
Principal investigator/supervisor	
Extension to approval needed (extensions are given for one year)	
Other	

Details of modification (give details of each of the amendments requested, state where the changes have been made and attach all amended and new documentation)

Research Procedure: The study experiment will occur in-person in the City research laboratory or through Mechanical Turk online testing. Both locations will require completion of the study on Qualtrics software. Participants will be presented with an audio statement and asked to complete follow up questions in survey format (i.e. multiple choice and text entry boxes).

Participation group: Participants will be 1st year Psychology students and public participants recruited on Amazon's Mechanical Turk for compensation.

Revised incentive: Participants will receive course credit or monetary compensation for taking part in the study.

Participant recruitment: Student participants will be recruited through City SONA online recruitment database. Public participants will be recruited through the Amazon Mechanical Turk marketplace.

Consent form/Information sheet: Both the consent form and information sheet will be included in electronic format in the Qualtrics software. Participants must provide consent to move forward to the information sheet and must read the information sheet before proceeding with the experiment.

Justify why the amendment/extension is needed (including the period of extension being requested)

Changes were made to the research software and procedure, participant group, participant incentive and recruitment.

Period of extension requested

Other information (provide any other information which you believe should be taken into account during ethical review of the proposed changes)

The study can be retrieved on this URL:
https://cityunilondon.eu.qualtrics.com/jfe/form/SV_1TgYOpuYni72S7X

Change in the study team

Staff member

<i>Title, Name & Staff Number</i>	<i>Post</i>	<i>Dept & School</i>	<i>Phone</i>	<i>Email</i>	<i>Date and type of CRB disclosure*</i>

Student

<i>Name & Student Number</i>	<i>Course / Year</i>	<i>Dept & School</i>	<i>Date and type of CRB disclosure*</i>

External co-investigator/s

<i>Title & Name</i>	<i>Post</i>	<i>Institution</i>	<i>Phone</i>	<i>Email</i>	<i>Date and type of CRB disclosure*</i>

Declaration (to be signed by the Principal Investigator)

- I certify that to the best of my knowledge the information given above, together with any accompanying information, is complete and correct and I take full responsibility for it.

Principal Investigator(s) (student and supervisor if student project)		
Date	8/5/2018	

Reviewer signature To be completed upon FINAL approval of the amendment.		
	Signature (Please type name)	Date

References

- Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, 49(1), 97–110.
<https://doi.org/10.3758/s13428-015-0689-6>
- Areh, I. (2011). Gender-related differences in eyewitness testimony. *Personality and Individual Differences*, 50, 559–563. <https://doi.org/10.1016/j.paid.2010.11.027>
- Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, 225(2), 82–90.
<https://doi.org/10.1038/scientificamerican0871-82>
- Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Exploring Working Memory: Selected Works of Alan Baddeley*, 4(11), 417–423. <https://doi.org/10.4324/9781315111261>
- Baddeley, A. D. (2003). WORKING MEMORY: LOOKING BACK AND LOOKING FORWARD. *Nature Reviews Neuroscience*, 4(829–839).
<https://doi.org/10.1038/nrn1201>
- Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation*, 8, 47-89. doi:10.1016/S0079-7421(08)60452-1
- Baddeley, A.D., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language*, 27(5), 586–595. [https://doi.org/10.1016/0749-596X\(88\)90028-9](https://doi.org/10.1016/0749-596X(88)90028-9)
- BBC. (n.d.). Desert Island Discs.
- BBC. (2018, November 22). BBC London News - Self Evident App Suspended
- Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). The effect of talker, rate, and amplitude variation on memory representation of spoken words. *Perception*

- & *Psychophysics*, 61(2), 206–219. <https://doi.org/10.1121/1.415243>
- Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review*, 100(1), 42–67. <https://doi.org/10.1037/0033-295X.100.1.42>
- Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology*, 15(1), 77–96. <https://doi.org/10.1348/135532509X414765>
- Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the Bad Guy in a Lineup Using Confidence Judgments Under Deadline Pressure. *Psychological Science*, 23(10), 1208–1214. <https://doi.org/10.1177/0956797612441217>
- Campeanu, S., Craik, F. I. M., & Alain, C. (2015). Speaker's voice as a memory cue. *International Journal of Psychophysiology*, 95(2), 167–174. <https://doi.org/10.1016/j.ijpsycho.2014.08.988>
- Campos, L., & Alonso-Quecuty, M. L. (2006). Remembering a criminal conversation: Beyond eyewitness testimony. *Memory*, 14(1), 27–36. <https://doi.org/10.1080/09658210444000476>
- Chan, E., Paterson, H. M., & van Golde, C. (2019). The effects of repeatedly recalling a traumatic event on eyewitness memory and suggestibility. *Memory*, 27(4), 536–547. <https://doi.org/10.1080/09658211.2018.1533563>
- Clifford, B. R. (1980). Voice Identification by Human Listeners: On Earwitness Reliability. *Law and Human Behavior*, 4(4), 373–393.
- Clifford, B. R., Rathborn, H., and Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior*, 5, 201–208.
- Cook, S., & Wilding, J. (1997a). Earwitness Testimony: Never Mind the Variety,

- Hear the Length. *Applied Cognitive Psychology*, 11(2), 95–111.
[https://doi.org/10.1002/\(SICI\)1099-0720\(199704\)11:2<95::AID-ACP429>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1099-0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O)
- Cook, S., & Wilding, J. (1997b). Earwitness Testimony 2: Voices, Faces and Context. *Applied Cognitive Psychology*, 11(6), 527–541.
[https://doi.org/10.1002/\(SICI\)1099-0720\(199712\)11:6<527::AID-ACP483>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1099-0720(199712)11:6<527::AID-ACP483>3.0.CO;2-B)
- Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, 96, 341–370.
- Deffenbacher, K. A., Cross, J. F., Handkins, R. E., Chance, J. E., Goldstein, A. G., Hammersley, R., & Read, J. D. (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *Journal of Psychology: Interdisciplinary and Applied*, 123(2), 109–119.
<https://doi.org/10.1080/00223980.1989.10542967>
- Dewhurst, S. A., Bould, E., Knott, L. M., & Thorley, C. (2009). The roles of encoding and retrieval processes in associative and categorical memory illusions. *Journal of Memory and Language*, 60, 154–164.
<https://doi.org/10.1016/j.jml.2008.09.002>
- Dewhurst, S. A., & Knott, L. M. (2010). Investigating the encoding–retrieval match in recognition memory: Effects of experimental design, specificity, and retention interval. *Memory & Cognition*, 38(8), 1101–1109.
<https://doi.org/10.3758/MC.38.8.1101>
- Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *American Journal of Psychology*, 111(2), 191–214.
<https://doi.org/10.2307/1423486>

- Dutton, A. & Carroll, M. (2001). Eyewitness testimony. *Australian Journal of Psychology*, 53(2), 83-91. <https://doi.org/10.1080/00049530108255128>
- Fueller, C., Loescher, J., & Indefrey, P. (2013). Writing superiority in cued recall. *Frontiers in Psychology*, 4, 1-12. doi: 10.3389/fpsyg.2013.00764.
- Goldinger, S. D. (1996). Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1166–1183.
- Goldstein, A. G., Chance, J. E., & Schneller, G. R. (1989). Frequency of eyewitness identification in criminal cases: A survey of prosecutors. *Bulletin of the Psychonomic Society*, 27(1), 71–74. <https://doi.org/10.3758/BF03329902>
- Grabowski, J. (2005). The written superiority effect. *Journal of Psychology*, 213, 193–204. doi: 10.1026/0044-3409.213.4.193
- Heath, A. J., & Moore, K. (2011). *Earwitness Memory : Effects of Facial Concealment on the Face Overshadowing Effect*. 33(February 2017), 131–140.
- Hollien, H. (2012). ON EARWITNESS LINEUPS. *Investigative Sciences Journal*, 4(1), 1–17. Retrieved from www.InvestigativeSciencesJournal.org
- Holt, N. E., Smith, R. E., Bremner, A. E., Sutherland, E. E., Vliek, M. E., & Passer, M. E. (2019). *Psychology: The Science of Mind and Behaviour* (4th ed.). London: McGraw-Hill Education.
- Home Office (2003) Advice on the Use of Voice Identification Parades. Circular 057/2003. London: *Home Office*.
- Horgan, T. G., Mast, M. S., Hall, J. A., & Carter, J. D. (2004). Gender Differences in Memory for the Appearance of Others. *Personality and Social Psychology Bulletin*, 30(2), 185–196. <https://doi.org/10.1177/0146167203259928>
- Howe, M. L., Knott, L. M., & Conway, M. A. (2017). Memory and miscarriages of

- justice. In *Memory and Miscarriages of Justice*.
- <https://doi.org/10.4324/9781315752181>
- Jacoby, L. L. (2010). Memory observed and memory unobserved. In *Remembering reconsidered* (pp. 145–177). <https://doi.org/10.1017/cbo9780511664014.007>
- Jaroslawska, A. J., Gathercole, S. E., & Holmes, J. (2018). Following instructions in a dual-task paradigm: Evidence for a temporary motor store in working memory. *Quarterly Journal of Experimental Psychology*, 71(11), 2439–2449. <https://doi.org/10.1177/1747021817743492>
- Johnson, M. K. (1997). Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society B*, 352, 1733–1745.
- Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, 20(2), 187–197. <https://doi.org/10.1002/acp.1175>
- Kerstholt, J. H., Jansen, N. J. M., Van Amersvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, 18(3), 327–336. <https://doi.org/10.1002/acp.974>
- Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and Sequential Lineups: Decision Processes of Accurate and Inaccurate Eyewitnesses. *Applied Cognitive Psychology*, 15(6), 659–671. <https://doi.org/10.1002/acp.739>
- Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA- Working Papers in Phonetics*, 49, 43–51.
- Laub, C. E., Wylie, L. E., & Bomstein, B. H. (2013). CAN THE COURTS TELL AN EAR FROM AN EYE? LEGAL APPROACHES TO VOICE IDENTIFICATION EVIDENCE. *Law & Psychology Review*, 37, 119–158.

- Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology (2006)*, 72(9), 2240–2248. <https://doi.org/10.1177/1747021819836890>
- Leander, L., Granhag, P. A., & Christianson, S. Å. (2005). Children exposed to obscene phone calls: What they remember and tell. *Child Abuse & Neglect*, 29(8), 871–888. <https://doi.org/10.1016/j.chiabu.2004.12.012>
- Legge, G. E., Grossman, C., & Pieper, C.M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2), 298-303.
- Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A. M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology*, 14(1), 13–24. <https://doi.org/10.1348/135532508X382708>
- Loftus, E. F. (1975). RECONSTRUCTING MEMORY: THE INCREDIBLE EYEWITNESS*. *Jurimetrics Journal*, 15(3), 188–193. Retrieved from <https://about.jstor.org/terms>
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Cold Spring Harbor Laboratory Press*, 361–366. <https://doi.org/10.1101/lm.94705.recalled>
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5), 585–589. [https://doi.org/10.1016/S0022-5371\(74\)80011-3](https://doi.org/10.1016/S0022-5371(74)80011-3)
- Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup

- instructions and the absence of the offender. *Journal of Applied Psychology*, 66(4), 482–489. <https://doi.org/10.1037/0021-9010.66.4.482>
- Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9, 111–151. [https://doi.org/10.1016/0010-0285\(77\)90006-8](https://doi.org/10.1016/0010-0285(77)90006-8)
- Mansour, J. K., Beaudry, J. L., Bertrand, M. I., Kalmet, N., Melsom, E. I., & Lindsay, R. C. L. (2012). Impact of disguise on identification decisions and confidence with simultaneous and sequential lineups. *Law and Human Behavior*, 36(6), 513–526. <https://doi.org/10.1037/h0093937>
- Manzanero, A. L., & Barón, S. (2017). Recognition and discrimination of unfamiliar male and female voices. *Behavior & Law Journal*, 3(1), 52–60.
- Markham, D. and Hazan, V. (2002). The UCL Speaker Database. Speech, Hearing and Language: UCL Work in Progress, vol. 14, p.1-17.
- Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of Talker Variability on Recall of Spoken Word Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 676–684. <https://doi.org/10.1121/1.397688>
- McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology*, 17(2), 249–271. <https://doi.org/10.1080/00221309.1937.9917999>
- McGehee, F. (1944). An experimental study of voice recognition. *The Journal of General Psychology*. 31. 53-65. doi:10.1080/00221309.1944.10545219
- McGorrery, P. G., & McMahon, M. (2017). A fair ‘hearing.’ *The International Journal of Evidence & Proof*, 21(3), 262–286. <https://doi.org/10.1177/1365712717690753>

- Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 2, 1–21.
[https://doi.org/10.1016/S0022-5371\(63\)80063-8](https://doi.org/10.1016/S0022-5371(63)80063-8)
- Memon, A., Havard, C., Clifford, B., Gabbert, F., & Watt, M. (2011). A field evaluation of the VIPER system: A new technique for eliciting eyewitness identification evidence. *Psychology, Crime, & Law*, 17(8), 711–729.
- Mickes, L., & Wixted, J. T. (2015). On the Applied Implications of the “Verbal Overshadowing Effect.” *Perspectives on Psychological Science*.
<https://doi.org/10.1177/1745691615576762>
- Mori, K., & Kishikawa, T. (2014). Co-Witness Auditory Memory Conformity following Discussion: A Misinformation Paradigm. *Perceptual and Motor Skills*, 118(2), 533–547. <https://doi.org/10.2466/24.22.pms.118k22w4>
- Mullennix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., ... Tessmer, B. (2010). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology*, 24(4), 513–526.
<https://doi.org/10.1002/acp.1566>
- Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21(4), 422–432. <https://doi.org/10.1177/026192702237958>
- Neisser, U. (1981). Dean white house verbatim memory. *Cognition*, 9(1), 1–22.
[https://doi.org/10.1016/0010-0277\(81\)90011-1](https://doi.org/10.1016/0010-0277(81)90011-1)
- Nevid, J. S. (2013). *Psychology: concepts and applications*. Belmont, CA: Wadsworth Cengage Learning.
- Nolan, F. (1997). *Speaker recognition and forensic phonetics*. In: W. J. Hardcastle & J. Laver (eds.) *The Handbook of Phonetic Sciences*, Oxford: Blackwell. pp.

744-767.

- Öhman, L., Eriksson, A., & Granhag, P. A. (2010). Mobile phone quality vs. Direct quality: How the presentation format affects earwitness identification accuracy. *European Journal of Psychology Applied to Legal Context*, 2(2), 161–182.
- Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the Planning of A Crime: Do Adults Outperform Children As Earwitnesses? *Journal of Police and Criminal Psychology*, 26(2), 118–127. <https://doi.org/10.1007/s11896-010-9076-5>
- Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Enhancing Adults' and Children's Earwitness Memory: Examining Three Types of Interviews. *Psychiatry, Psychology and Law*, 20(2), 216–229. <https://doi.org/10.1080/13218719.2012.658205>
- Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, 9(3), 249–260. <https://doi.org/10.1002/acp.2350090306>
- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2), 309–328. https://doi.org/10.1007/978-3-662-13015-5_21
- Paterson, H. M., van Golde, C., Devery, C., Cowdery, N., & Kemp, R. (2018). iWitnessed: Capturing Contemporaneous Accounts to Enhance Witness Evidence. *Current Issues in Criminal Justice*, 29(3), 273–281. <https://doi.org/10.1080/10345329.2018.12036102>
- Penrod, S., & Cutler, B. (1995). Witness Confidence and Witness Accuracy: Assessing Their Forensic Relation. *Psychology, Public Policy, and Law*, 1(4),

817–845. <https://doi.org/10.1037/1076-8971.1.4.817>

Perfect, T. J., Hunt, L. J., & Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology*. <https://doi.org/10.1002/acp.920>

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, 3(FEB), 1–11.
<https://doi.org/10.3389/fpsyg.2012.00023>

Pezdek, K., & Prull, M. (1993). Fallacies in memory for conversations: Reflections on Clarence Thomas, Anita Hill, and the Like. *Applied Cognitive Psychology*, 7, 299–310.

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, 21(4), 539–550.
<https://doi.org/10.1002/acp.1296>

Pickel, K. L., & Staller, J. B. (2012). A perpetrator's accent impairs witnesses' memory for physical appearance. *Law and Human Behavior*, 36(2), 140–150.
<https://doi.org/10.1037/h0093968>

Possley, M. (2018, October 13). Wilber Jones. Retrieved from
<https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=5392>.

Prisonstudies.org. (2019a). United Kingdom: England & Wales | World Prison Brief. [online] Available at: <https://www.prisonstudies.org/country/united-kingdom-england-wales> [Accessed 10 Nov. 2019].

Prisonstudies.org. (2019b). United States of America | World Prison Brief. [online] Available at: <https://www.prisonstudies.org/country/united-states-america> [Accessed 10 Nov. 2019].

R v Nealon in EWCA Crim 574. 2014.

Read, D., & Craik, F. I. M. (1995). Earwitness Identification: Some Influences on Voice Recognition. *Journal of Experimental Psychology: Applied*, 1(1), 6–18.
<https://doi.org/10.1037/1076-898X.1.1.6>

Rechdan, J., Sauer, J. D., Hope, L., Sauerland, M., Ost, J., & Merckelbach, H. (2017). Computer mediated social comparative feedback does not affect metacognitive regulation of memory reports. *Frontiers in Psychology*, 8(AUG), 1–11. <https://doi.org/10.3389/fpsyg.2017.01433>

Reisberg, D. (2014). The Science of Perception and Memory. In *The Science of Perception and Memory: A Pragmatic Guide for the Justice System Daniel*.
<https://doi.org/10.1093/acprof:oso/9780199826964.001.0001>

Robson, J. (2017). A fair hearing? The use of voice identification parades in criminal investigations in England and Wales. *Criminal Law Review*, 1, 36–50.
<https://doi.org/10.1177/1365712718782989>

Robson, J. (2018). ‘Lend me your ears’: An analysis of how voice identification evidence is treated in four neighbouring criminal justice systems. *International Journal of Evidence and Proof*, 22(3), 218–238.
<https://doi.org/10.1177/1365712718782989>

Roebuck, R. & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology*, 7, 475–481.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, 16(2), 252–257. <https://doi.org/10.3758/PBR.16.2.252>

Saltzburg, Stephen A., Voice Identification Experts (2013). 27 *Crim. Just.* (2013);

GWU Law School Public Law Research Paper No. 2013-143; GWU Legal

Studies Research Paper No. 2013-143. Available at SSRN:

<http://ssrn.com/abstract=2663669>.

Sanders, G. & Warnick, D. (1980). Some conditions maximizing eyewitness accuracy: A learning/memory analogy. *Journal of Criminal Justice*, 8, 395-403. 10.1016/0047-2352(80)90115-4.

Sarwar, F., Allwood, C. M., & Zetterholm, E. (2014). Earwitnesses: The type of voice lineup affects the proportion of correct identifications and the realism in confidence judgments. *International Journal of Speech, Language and the Law*, 21(1), 139–155. <https://doi.org/10.1558/ijssl.v21i1.139>

Saslove, H., & Yarmey, A. D. (1980). Long-Term Auditory Memory: Speaker Identification. *Journal of Applied Psychology*, 65(1), 111–116.

Sauerland, M., & Sporer, S. L. (2011). Written vs. spoken eyewitness accounts: does modality of testing matter?. *Behavioral Sciences & The Law*, 29(6), 846–857. <https://doi.org/10.1002/bsl.1013>

Sauerland, M., Krix, A. C., van Kan, N., Glunz, S., & Sak, A. (2014). Speaking is silver, writing is golden? The role of cognitive and social factors in written versus spoken witness accounts. *Memory & Cognition*, 42(6), 978–992. <https://doi.org/10.3758/s13421-014-0401-6>

Schneider, W. (2015). Memory development from early childhood through emerging adulthood. In *Memory Development from Early Childhood Through Emerging Adulthood*. <https://doi.org/10.1007/978-3-319-09611-7>

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal Overshadowing of Visual Memories: Some Things Are better Left Unsaid Experiments 1-5 were included in. *Cognitive Psychology*, 22, 36–71.

- Semmler, C., Brewer, N., & Douglass, A. B. (2012). Jurors believe eyewitnesses. In B. L. Cutler (Ed.), *Conviction of the innocent: Lessons from psychological research* (p. 185–209). American Psychological Association.
<https://doi.org/10.1037/13085-009>
- Sherrin, C. (2015). Earwitness Evidence: The Reliability of Voice Identifications. *Osgoode Hall Law Journal*, 52(3), 1–44. <https://doi.org/10.2139/ssrn.2628313>
- Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, 296, 131–140.
<https://doi.org/10.1016/j.heares.2012.11.004>
- Smith, H. M.J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: optimising witness performance. *Memory*, 28(1), 2–17. <https://doi.org/10.1080/09658211.2019.1673427>
- Smith, H. M. J., & Baguley, T. (2014). Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students*, 5(1), 59–68. <https://doi.org/10.5334/jeps.bs>
- Spence, K., Arciuli, J., & Villar, G. (2012). *The role of pitch and speech rate as markers of deception in Italian speech*, presented at Australasian International Conference on Speech Science and Technology, Sydney, 2012. Sydney, Australia: Macquarie University.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. In *Behavior Research Methods, Instruments, & Computers*.
- State v. Hauptmann*, Atlantic Rep., 1935, 180, 809-829.
- Stebay, N. M. (1997). Social Influence in Eyewitness Recall: A Meta-Analytic Review of Lineup Instruction Effects. In *Source: Law and Human Behavior* (Vol. 21).

- Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology*, 24(6), 647–653.
<https://doi.org/10.1080/20445911.2012.675321>
- Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, 25(1), 112–118.
<https://doi.org/10.1002/acp.1649>
- The Innocence Project. (n.d.). <https://www.innocenceproject.org/>
- The Smart Way To Report Crime. Just Evidence, Provided by the Self Evident App.
Retrieved December 20, 2019, from <https://www.justevidence.org/>
- Tulving, E., & Thomson, D. M. (1973). ENCODING SPECIFICITY AND RETRIEVAL PROCESSES IN EPISODIC MEMORY. *Psychological Review*, 80(5), 352–373.
- Vanags, T., Carroll, M., & Perfect, T. J. (2005). Verbal overshadowing: A sound theory in voice recognition? *Applied Cognitive Psychology*, 19(9), 1127–1144.
<https://doi.org/10.1002/acp.1160>
- Vanden Abeele, M. M. P. (2016). Mobile youth culture: A conceptual development. *Mobile Media and Communication*, 4(1), 85–101.
<https://doi.org/10.1177/2050157915601455>
- Vatanasuk, N., Chomputawat, A., Chomputawat, S., & Chatwiriya, W. (2015). Mobile Crime Incident Reporting System using UX dimensions guideline. *ACDT 2015 - Proceedings: The 1st Asian Conference on Defence Technology*, 187–192. <https://doi.org/10.1109/ACDT.2015.7111609>
- Wells, G. L. (2001). Police Lineups: Data, Theory, and Policy. *Psychology, Public Policy, and Law*, 7(4), 791–801. <https://doi.org/10.1037/1076-8971.7.4.791>
- Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., &

- Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, 44(1), 3-36. <http://dx.doi.org/10.1037/lhb0000359>
- Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness Identification Procedures: Recommendations for Lineups and Photospreads. *Law and Human Behavior*, 22(6), 1-39.
- Wilding, J. & Cook, S. (2000). Sex differences and individual consistency in voice identification. *Perceptual and Motor Skills*, 91, 535–538.
- Winkler, I., & Cowan, N. (2005). From sensory to long-term memory: Evidence from auditory memory reactivation studies. *Experimental Psychology*, 52(1), 3–20. <https://doi.org/10.1027/1618-3169.52.1.3>
- Wixted, J. T., Don Read, J., & Stephen Lindsay, D. (2016). The Effect of Retention Interval on the Eyewitness Identification Confidence–Accuracy Relationship. *Journal of Applied Research in Memory and Cognition*, 5(2), 192–203. <https://doi.org/10.1016/j.jarmac.2016.04.006>
- Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica*, 114, 101–114. [https://doi.org/10.1016/S0001-6918\(03\)00052-0](https://doi.org/10.1016/S0001-6918(03)00052-0)
- Yarmey, A. D. (1991). Voice Identification Over the Telephone. *Journal of Applied Social Psychology*, 21(22), 1868–1876. <https://doi.org/10.1111/j.1559-1816.1991.tb00510.x>
- Yarmey, A. D. (1995). Earwitness speaker identification. Special Issue: Witness memory and law. *Psychology, Public Policy, and Law*, 1(4), 792–816.
- Yarmey, A. D. (2007). The psychology of speaker identification and earwitness memory. In *The Handbook of Eyewitness Psychology: Volume II: Memory for*

- People* (Vol. 2, pp. 101–136). <https://doi.org/10.4324/9780203936368-11>
- Yarmey, A. D. (2012). Factors Affecting Lay Persons' Identification Of Speakers. In *The Oxford Handbook of Language and Law* (pp. 1–12). <https://doi.org/10.1093/oxfordhb/9780199572120.013.0040>
- Yarmey, A. D., & Matthys, E. (1992). Voice ID of an abductor. *Applied Cognitive Psychology*, 6, 367–377.
- Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense Beliefs and the Identification of Familiar Voices. *Applied Cognitive Psychology*, 15(3), 283–299. <https://doi.org/10.1002/acp.702>
- Zaragoza, M. S., Belli, R. F., & Payment, K. E. (2013). Misinformation Effects and the Suggestibility of Eyewitness Memory. In M. Garry & H. Hayne (Eds.), *Do Justice and Let the Sky Fall Elizabeth F. Loftus and her Contributions to Science, Law, and Academic Freedom* (pp. 35–63). MAHWAH, NEW JERSEY: LAWRENCE ERLBAUM ASSOCIATES.
- Zetterholm, E., Sarwar, F., Thorvaldsson, V., & Allwood, C. M. (2012). Earwitnesses: The effect of type of vocal differences on correct identification and confidence accuracy. *International Journal of Speech, Language and the Law*, 19(2), 219–237. <https://doi.org/10.1558/ijssl.v19i2.219>