Münsterberg’s Legacy: What Does Eyewitness Research Tell Us About the Reliability of Eyewitness Testimony?

AMINA MEMON*, SERENA MASTROBERARDINO and JOANNE FRASER

School of Psychology, University of Aberdeen, Scotland, UK

SUMMARY
Münsterberg alerted us about the fallibility of eyewitness memory 100 years ago and we have come far in our understanding of the factors contributing to eyewitness error. A healthy empirical literature, including several meta-analytic reviews, suggests progress has been made in understanding the impact of various estimator and system variables on eyewitness performance. We begin by acknowledging Münsterberg’s pioneering studies of eyewitness fallibility. We then turn to the methodological contribution of Buckhout (1974), a follower of Münsterberg. A selective review of research then highlights the scientific merits and pitfalls of the laboratory and archival approach to studying eyewitness behaviour. We end with a discussion of what experts can really tell the courts about the causal and associative relationships between various estimator and system variables.

INTRODUCTION
Over a century has passed since Hugo Münsterberg questioned the reliability of eyewitness testimony. His controversial book ‘On the Witness Stand’ (1908), was one of the earliest influences in the field of Psychology and Law¹. In the chapters on ‘illusions’ and ‘memory of the witness’ Münsterberg described his experimental demonstrations on the fallibility of eyewitness memory; however his work was met with fierce criticism (see Dolyle, 2005 for an account of John Henry Wigmore’s devastating legal critique). The unfashionable nature of Münsterberg’s research and the lack of any underlying explanations also fuelled criticisms from his colleagues. Interestingly, modern day eyewitness researchers have also stressed the need for theory driven research (see Clark, 2008). In the current paper, we draw analogies between some of the findings from Münsterberg’s early studies and the conclusions from recent laboratory and archival studies. We discuss the extent to which we can draw meaningful conclusions from laboratory research. In the final section we consider whether an understanding of the factors impacting eyewitness reliability enables ‘experts’ to predict the accuracy of real life eyewitnesses.

There are several strands of Münsterberg’s work that have had a significant impact on contemporary eyewitness research. Perhaps most relevant to eyewitnessing are

*Correspondence to: Amina Memon, School of Psychology, William Guild Building, University of Aberdeen, King’s College, Aberdeen AB24 2UB, Scotland, UK. E-mail: amemon@abdn.ac.uk

¹The contribution of Münsterberg’s contemporaries and predecessors should also be noted (see Bornstein & Penrod, 2008).

Copyright © 2008 John Wiley & Sons, Ltd.
Münsterberg’s scientific demonstrations of the limitations of eyewitness memory. He describes one such demonstration, in which children and adults viewed pictures of a farmer’s room and immediately afterwards individually answered a mixture of objective and leading questions (Münsterberg, 1908, p. 181). These studies showed leading questions can mislead witnesses and that younger people may be less resistant to suggestion than adults. Contemporary researchers continue to study the impact of misleading questions on children and adults (e.g., Roebers & Schneider, 2000). This research has improved our understanding of the impact of post-event information on memory and the cognitive, social and neurobiological mechanisms underlying memory distortions (See Loftus, 2005, for a review). The research on the impact of leading questions has had a significant impact on policy and practice, particularly the questioning of vulnerable witnesses (children, those with a disability and intimidated witnesses). Examples from the UK include the Vulnerable Witness (Scotland) Act, 2004 and Achieving Best Evidence (http://www.homeoffice.gov.uk/documents/achieving-best-evidence). Thus in the UK there is extensive guidance for professionals in the planning and conduct of interviews as well as information on witness support and preparation for the court process.

As an illustration of the discrepancies between observers in their perception of an event, Münsterberg presented psychology students with a piece of white cardboard with 50 black squares on it. When asked how many squares there were, the answers ranged from 25 to 200. Only three students were able to give the correct answer. A repeat of the experiment with 20 black squares produced similar results. As pointed out by Münsterberg these participants were highly trained, careful observers whose attention was concentrated on the material (Münsterberg, 1908, cited by Dolyle, 2005). Münsterberg also provided a simple demonstration of attentional focus in which the experimenter performed a different task with each of his hands. With his right hand the experimenter was moving a little revolving wheel while the left was performing other actions (writing, opening a box, etc). Participants were asked to pay attention to what the experimenter was doing during a specified time period. When asked to report everything the experimenter had done, participants did not report any actions performed with the left hand. Later in this paper, we examine how the impact of attentional focus has been studied in contemporary studies of eyewitness memory.

A final example is Münsterberg’s presentation of a series of studies showing that a witness’s certainty has no bearing on the attention with which an object was observed. Here Münsterberg was alluding to the complex relationships among attention, recollection and the feeling of confidence (Dolyle, 2005). Once again in line with Münsterberg’s research, the conditions under which confidence may or may not be a marker of accuracy is a question that has occupied contemporary researchers in the eyewitness arena (see Brewer and Weber, 2008).

BOB BUCKHOUT TO THE RESCUE

Bob Buckhout, brandishing the fallen banner of Hugo Münsterberg, conducted raids against the criminal justice system’s complacent reliance on eyewitness testimony (Dolyle, 2005). Buckhout was on an explicit mission to rescue Münsterberg’s reputation and Münsterberg’s insights from ‘the Valley of Lost Things’ (Dolyle, 2005, p. 52).

Buckhout was the first to show how eyewitness conditions could be simulated on university campuses. In 1974, Buckhout staged a mock assault in front of 141 unsuspecting college students. Seven weeks later, these students were asked to pick the perpetrator from
a group of six photographs. Of the 60% who did not correctly identify the assailant, two-thirds incorrectly chose an innocent bystander who was at the crime scene. This innocent bystander effect has been replicated in numerous studies and has resulted in theoretically driven research aimed at identifying the cognitive and social mechanisms underlying the effect (Brainerd & Reyna, 2005; Deffenbacher et al., 2006).

Another early demonstration of the efficacy of the staged event scenario, this time focusing on system variables, is Malpass and Devine’s (1981) simulation study of the effects of biased lineup instructions on eyewitness identification ability. The researchers staged a crime in the middle of a lecture, during which a single perpetrator entered the lecture hall, exchanged heated words with the instructor, and knocked over a rack of machines. The witnesses made more or less accurate identifications from a photo lineup, depending on the instructions the researchers gave them. The differences between the two sets of instructions were subtle: one set implied that the witness had to identify the perpetrator from the lineup, conveying to the witness that he was present. The other set of instructions gave the witness the option of not making a choice, implying that the perpetrator may not be present.

Malpass and Devine (1981), in order to make it realistic, included the real perpetrator in the lineup only half the time. When witnesses thought they had to choose, they were more likely to pick the wrong suspect. When witnesses thought they did not have to choose, however, they rarely made false identifications. When making a correct identification witnesses who were not induced to choose were just as accurate as were subjects who were induced to choose. Malpass and Devine concluded that simply telling witnesses that they do not have to choose one of the suspects in the lineup causes them to make fewer false identifications without hindering witnesses’ ability to make accurate identifications. The effect of biased instructions has been replicated in numerous studies over the years (see Clark, 2005 for a recent review) and has impacted policy and practice relating to the administration of identification parades (e.g. Davies & Valentine, 1999; Wells et al., 2000).

To overcome some of the practical difficulties associated with staging crimes, researchers have relied extensively on the use of filmed presentations. In one of the earliest studies using a filmed presentation of a crime, Clifford and Hollin (1981) showed that the presence of multiple perpetrators substantially reduced identification performance. It is interesting to note that despite clear effects of multiple perpetrators on identification accuracy, eyewitness studies conventionally only have a single perpetrator. Moreover, many researchers share stimuli and findings are thus replicated with the exact same stimuli, target persons and lineups. This is not good practice if the aim is to apply the research findings to a wide range of eyewitnessing scenarios and targets varying in distinctiveness. Moreover, witnesses are typically presented with static views of faces at identification despite the fact that a dynamic image is presented at study. However, as a result of changes in procedures for obtaining identification evidence in the UK, researchers have begun to examine whether moving images at test can confer advantages (Darling, Valentine, & Memon, 2008; Valentine, Darling, & Memon, 2007).

THE TYPICAL WITNESS AND SCENARIO: ETHICAL AND PRACTICAL CONSTRAINTS

There is no disputing the fact that eyewitness simulation studies can only capture a limited set of the conditions that occur in real world situations, an issue we will return to later. Ethical constraints severely restrict what researchers can do and practical constraints have
resulted in an over-representation of college students in study samples. However, let us not overlook the fact that the literature does contain studies using more diverse samples from child witnesses and seniors, to adults and children with intellectual disabilities (see Memon, Vrij, & Bull, 2003; Toglia, Read, Ross, & Lindsay, 2006, Volume 1). Studies that have compared students to non-students show that in terms of absolute levels of recall and identification, students are no different from members of the public (Loftus, Levidow, & Duensing, 1992). Importantly, researchers have acknowledged that there are similarities as well as differences between eyewitnesses from different populations and age groups. For example, younger and older adults are equally susceptible to the innocent bystander effect despite age-related increases in false identifications (Memon, Hope, & Bull, 2003). On the other hand, preschoolers are more suggestible than adults and older children (Ceci & Bruck, 1993). The limited amount of research with police as eyewitnesses suggests they provide more correct descriptors than civilians but are no more reliable as eyewitnesses (Zimmerman, 2006). When it comes to judging the accuracy of eyewitness statements, recent research from Sweden suggests that police outperform laypersons in assessing the reliability of individual eyewitness statements (Lindholm, 2008).

In terms of sampling the range of situations that could occur in the real world, concerns about the welfare of human subjects has meant researchers have relied extensively on non-threatening scenarios where the witnesses are bystanders rather than victims. Nevertheless the staged event procedure and more recently the use of video-taped scenarios has enabled researchers to simulate a variety of witnessing conditions that are not that far removed from the kinds of experiences an individual may encounter as a victim or witness in the real world. For example, Stein and Memon (2006) recently simulated an abduction in which the victim was kidnapped, held at gunpoint, and robbed before being released by several perpetrators. The witnesses in the study were the cleaning staff of the university and they were interviewed with a procedure, the Cognitive Interview, that has been found to increase the number of correctly reported details (Köhneken, Milne, Memon, & Bull, 1999). We replicated a well established finding based almost exclusively on college student samples, namely that the Cognitive Interview yields significantly more correct details than the typical police interview. To increase the ecological validity of our study, our control interview was modelled upon the typical police procedure employed in Brazil. One limitation of the Stein and Memon (2006) study was the witnesses could only be subjected to stress using a videotaped simulation and hence the impact of personal threat on eyewitness performance could not be examined.

The ethical dilemmas associated with simulating stress have been tackled in a series of studies recently using citizens who experience high intensity stress as part of their job. Morgan et al. (2004) examined the identification abilities of military personnel who were subject to an interrogation as part of military survival training. The participants were exposed to the stress of food and sleep deprivation 48 hours prior to being subjected to interrogation stress. The findings are quite complex and subject to some methodological confounds but overall indicate impairments in face recognition accuracy under high stress conditions. Interestingly, there were no associations between demographic variables (age, rank, years of experience, trauma history) and identification accuracy. However, in a later study, Morgan et al. (2007) found participants’ ability to remember faces on a standardized face recognition test was significantly associated with the ability of the military personnel to recognise the face of the interrogator. These findings suggest that laboratory measures can have some external validity. Nevertheless laboratory studies alone do not present a complete picture of all the variables that can impact eyewitness performance. There are two
other methods that have been used: archival methods and field experiments. Archival studies and field experiments have been utilised in the quest for ecological validity. The advantages and disadvantages of these techniques are the focus of the next section.

ARCHIVAL AND FIELD STUDIES

Archival studies typically rely on police files of cases that have been prosecuted. Researchers study the statements of witnesses to examine the characteristics of witnesses and the completeness and accuracy of the information reported, against various official sources that contain information about the crime and perpetrator. Archival studies provide an opportunity to look at a complex set of variables (see Tollestrup, Turtle, & Yuille, 1994) that often cannot be manipulated in the laboratory and hence data from these sources cannot be directly compared. However, archival data could serve to enrich the data obtained in the laboratory. That said there are several shortcomings: identifying the accuracy of archival material, such as the accuracy of perpetrator descriptions, and corroborating the content of the report. This is because of the assumption that the suspect is indeed the perpetrator and that the suspect’s appearance has not changed substantially between the crime event and lineup. The prevalence of these sources of errors is virtually impossible to predict (van Koppen & Lochun, 1997). Another shortcoming of archival research is that errors of omission are typically unknown so it is virtually impossible to get an idea of how complete witness reports are. In order to give a flavour as to the complexity of the archival research method, several studies will be reviewed here.

Van Koppen and Lochun (1997) were given access to official court records kept by prosecution offices in the Netherlands for 1992. The case files included all statements recorded by police, findings of the police investigation, suspect and witness interrogations and written decisions of the court. The data set included all robberies for which one or more of the offenders were convicted. Most witness statements were recorded shortly after the robbery (same day, next day). A total of 431 robberies in which 1313 witnesses gave 2299 descriptions were included in the sample. The data on the perpetrators’ true appearance was extracted from the national database of offenders maintained by the police. A total of 582 robbers participated in the crimes; however the witness descriptions could be compared to the appearance of only 462 robbers due to limitations of the database. Independent variables comprised both robber and witness characteristics. Witness characteristics included sex, age, type of witness and relationship to the victim (e.g., passer-by). Robbery characteristics included lighting conditions, obstructions to the witness view (e.g., glass pane), estimated distance from robber, level of threat and facial disguise. The numbers of descriptor units mentioned in each description were counted, as were the number of permanent characteristics and temporary characteristics (e.g., clothing). Descriptions were scored as wrong, correct, and partially correct against the police file based upon agreement between coders. The most frequently mentioned characteristic was the sex of the suspect. Over half of the descriptions also contained information about age, appearance (including race), skin colour and type of head covering or disguise. Overall the descriptions contained more permanent than temporary characteristics and the completeness of descriptions was poor with an average of eight descriptors. Hair colour descriptions were correct in 73% of cases while reports of facial hair were almost always incorrect. The overall accuracy of descriptions mentioned was rather high at 72%. A shorter delay between the crime and provision of witness description
and a shorter distance between the witness and robbery were associated with more complete descriptions while duration of the incident was not associated with any of the measures. However, several limitations of the dataset used by Van Koppen and Lochan can be identified. For example, the case files contained no information about factors that could have influenced the quality of the witness statement such as how witnesses were interviewed, the race of the witness, the level of intoxication of the witness and whether the witness was wearing corrective lenses.

Valentine, Pickering, and Darling (2003) studied the identification attempts of a total of 640 witnesses from 314 lineups conducted in London. The investigating officers provided demographic details of the witnesses and information about viewing conditions, exposure, distance, lighting, witness role (bystander, friend, victim), whether a weapon was present, suspect physical characteristics and match of description to suspect, offence type and delay since witnessing the event. Information about the outcome of the lineup (suspect identification, foil identification, no identification) was obtained from the identification officer. Most lineups had multiple witnesses. The age of the witness had a significant effect on identification accuracy with a rate of 48% suspect identifications in the under 20-year-old group as compared to 28% in the 40-plus group, which is consistent with the results of another British archival study (Pike, Brace, & Kyman, 2002). A higher number of witnesses fell in the middle-aged group (40–59) as compared to the older-adult (60–80) age range that is more conventional in laboratory research focusing on older adults. This makes comparisons of the Valentine et al. dataset with the laboratory literature difficult. On the basis of laboratory findings, an increase in mistaken (foil) identifications would have been expected among the older (60- to 80-year-old) group (see Bartlett & Memon, 2007 for a review of the laboratory studies).

Witness race and witness role had no significant effects on the outcome of the lineups in the Valentine et al. (2003) study, while witness gender only had the effect of increasing foil identifications. In terms of viewing conditions, witnesses who had more than one minute to observe the suspect tended to make more accurate identifications, a finding that is consistent with laboratory findings (Memon, Hope, & Bull, 2003). Lighting quality, obstructions and viewing distance had no notable effects on identification accuracy. Weapons (mostly knives) had no effect on suspect identification rates. Sixty-six percent of witnesses who gave a description coded as ‘detailed’ identified the suspect as compared to 14% who gave briefer details. These findings contrast with those of a recent meta-analysis of 33 research studies including a total sample of 4278 participants which suggests small or non-significant correlations between measures of description accuracy and subsequent identification accuracy, particularly in studies using identification paradigms (Meissner, Sporer, & Susa, 2008). In their discussion of their findings, Valentine et al. (2003) acknowledge the limitations of their data set in making inferences about causal mechanisms. As they point out archival data can only provide information about associative relationships among a limited set of variables.

An alternative to the laboratory method and archival analysis is the field experiment, and this is possibly the most under-utilised method by eyewitness researchers. The typical procedure is for a confederate to randomly approach members of the public on the pretence of asking for directions or using some similar excuse. A few minutes later the experimenter approaches the citizen asking if they can recognise the person that spoke to them earlier. Yarmey (1993) examined the recognition ability of 651 citizens after a 15-second exposure to a target. Witnesses recalled about 60% of the physical characteristics (hair colour, style, weight, height, age, eye colour). In a follow-up study of 603 citizens (Yarmey & Yarmey,
1997), recall accuracy for the target person ranged from 29% for hair style to 83% for hair colour. Recall of clothing ranged from 40% to 88% for upper and lower clothing, respectively. In these ecologically valid studies witnesses were not expecting a memory test but unlike real world situations the encounters were relatively mundane.

There are two notable examples of the use of the field experiment method to address specific hypotheses. Dysart, Lindsay, McDonald, and Wicke (2002) examined the effect of alcohol on identification accuracy from showups (a photo of the suspect) and the effects of viewing a showup on identification accuracy from lineups viewed a week later. They were interested in the impact of myopic processes in a person identification situation where under the influence of alcohol witnesses may not encode sufficient details to enable them to discriminate a target’s face from other faces. Dysart et al. (2002) measured the blood alcohol level of people who were exposed to a target in a social setting (a bar) and who were then (immediately) presented with a show-up (a single photo that was not the target). Identification performance declined with increasing alcohol with more choices of the innocent foil supporting alcohol myopia theory. Unfortunately due to sampling restrictions a systematic test of the hypothesis was not possible in a more formal photo lineup situation that was administered via a website a week later. The descriptive data from the self-selected group of witnesses who participated indicated that alcohol was associated with an increase in false identifications in target absent lineups.

Wright, Boyd, and Tredoux (2001) conducted a field study of the cross-race effect in South Africa and England. In each country, they had a black or white male confederate approach either a black or white member of the public in a shopping centre to enquire if they had seen some jewellery that they had lost. A few minutes later, another confederate asked questions about the stranger and asked if they could identify him from a lineup. An own-race bias was found in response to questions about the target as well as the ability to accurately identify the target from the lineup, supporting the findings of previous laboratory studies (Meissner & Brigham, 2001) as well as evidence from some archival research (Behrman & Davey, 2005).

**CAN WE RELY ON EXPERT TESTIMONY ON EYEWITNESS ISSUES?**

It was Münsterberg’s belief in the validity of experimental methods for the study of real world problems that led him to propose that experimental psychologists bring their work into the courtroom. Now that there is a large body of empirical research on the factors governing eyewitness behaviour the question arises as to whether or not experts agree on the findings.

One criticism that has been directed at experts is whether the information derived from their laboratory research based on specific groups, and expressed in terms of differences between groups and/or witnessing conditions can be meaningfully applied to a single case. The aim of expert testimony, however, is not to predict behaviour from a single case but to identify a set of variables or conditions that are associated with accurate or inaccurate eyewitness performance. It is then left up to the courts to decide if the evidence is relevant and how to apply an expert’s evidence in a specific case. Importantly, experts who routinely provide testimony do generally agree to the main findings as they pertain to the effects of various witness characteristics, witnessing conditions and retrieval contexts on the accuracy of an eyewitness’s report (Kassin, Tubb, Hosch, & Memon, 2001). Moreover, meta-analytic reviews of the literature have enabled some precise predictions to be made.
(see Deffenbacher, 2008). There may, however, be contradictions among experts in how the ‘data’ are explained to the courts. Wright (2006) highlights the importance of making a distinction between causal and associative hypotheses in psychology. Taking a topic already discussed in the current paper as an example, a causal hypothesis about the effect of stress on memory would state that a negative emotional event will adversely affect eyewitness accuracy and completeness. An expert may qualify this statement by saying it is based on the sample of events included in the Deffenbacher et al. (2004) meta-analysis. However, if an expert is asked are emotional events remembered more accurately than non-emotional events the answer would involve comparing emotional and non-emotional events. In other words this is a test of an associative hypothesis. As pointed out by Wright (2006) stressful events tend to differ from other events in many ways other than emotionality, so it is often difficult to isolate stress and assess causality. In an attempt to measure causal hypotheses about the effects of stress on memory, researchers have largely focused on varying the emotional nature of the stimuli such as the addition of an emotionally arousing slide in a sequence of neutral slides.

There is a problem in interpreting the effects of ‘emotional’ stimuli on memory. Are the effects due to the emotional nature of the slide or some kind of novelty or unusualness effect? Thus a stick of celery can produce an effect similar to holding a handgun in a threatening manner (Mitchell, Livosky, & Mather, 1998) and the weapon focus effect is more likely to occur where it is out of context (at a baseball match) as compared to an expected context (e.g., in a shooting range, Pickel (1999) or during police firearms training, Hulse & Memon, 2006). One explanation of the weapon focus effect is that unusual or unexpected objects drain our attentional resources resulting in poor memory for other aspects of the scene including the target irrespective of the level of threat and emotional arousal (see Hope & Wright, 2007). However, recent research has gone one step further in showing that the direction of visual attention to a weapon is not an involuntary or unavoidable response. Witnesses who are trained to do so will focus their attention elsewhere (Pickel, Ross, & Truelove, 2006). Thus when it comes to applying the findings of eyewitness research psychologists will be drawing attention to both causal and associative hypotheses and Wright’s (2006) analysis highlights the importance of this distinction. In doing so, Wright (2006) suggests that some of the experts sampled by Kassin et al. (2001) may have misinterpreted some of the associative relationships between variables on the survey as causal statements. The extent to which this overshadows the clear consensus among experts on a number of eyewitness issues is debatable. Needless to say Wright (2006) has drawn attention to the importance of presenting eyewitness facts based on associative versus causal hypotheses between various system and estimator variables appropriately in the courtroom.

CONCLUSION

More than 100 years have passed since Münsterberg first criticised the reliability of eyewitness testimony used in court. Though Münsterberg’s efforts were stalled, the significance of his early demonstrations is evident and parallels can be made with contemporary eyewitness research. Inspired by Münsterberg, Buckhout demonstrated how we could simulate eyewitnessing conditions in controlled settings. Since the 1970s we have used a variety of methods to study eyewitnesses and over the last 30 years have developed a good understanding of the conditions that lead to inaccuracies as well as ways of improving
accuracy (see Wells, Memon, & Penrod, 2006 for a review). In recent years, research efforts have been fuelled by analysis of DNA exonerations where wrongful convictions resulted from faulty eyewitness evidence. Moreover, there seems to be a reasonable degree of consensus as to the factors influencing eyewitness performance among experts who have been surveyed. As our knowledge base and methodological sophistication continue to expand, it is essential that we develop more precise theories to explain the variations in eyewitness performance that arise from complex interactions between variables.

Kassin, Ellsworth, and Smith (1989) and Kassin et al. (2001) surveys of eyewitness experts highlight the diversity of the eyewitness literature and the areas that have achieved the most focus. They also point out what areas, in the experts’ view, have achieved sufficient reliability that would lead these experts to testify in court. As pointed out by Wright (2006), the distinction between associative and causal hypotheses is an important one, especially when providing expert testimony in court. Researchers in this field should make clear whether they are referring to causal or associative relationships between variables. The challenge for future researchers is to develop innovative methods for simulating eyewitness behaviour (see Clark) as well as more precise methods for recording the accuracy of observations based on archival analysis. However, perhaps most important is to be able to add to, and modify existing theories about eyewitness behaviour, thereby ensuring that eyewitness research is valued by the legal system as Münsterberg intended it to be.

ACKNOWLEDGEMENTS

This paper was supported by a research grant from the European Union (Framework 6) awarded to the first author.

REFERENCES


Copyright © 2008 John Wiley & Sons, Ltd.


DOI: 10.1002/acp


