Court Relevant Research Methods

Matthew A. Palmer, Ruth Horry, & Neil Brewer

School of Psychology

Flinders University
COURT-RELEVANT RESEARCH METHODS

Background

Consider some of the key events that typically occur during a criminal court case. The jury will hear evidence relating to the events surrounding the crime in question. This evidence will usually be based, at least in part, on the recollections of one or more witnesses interviewed by police during their investigations. In many cases, the jury will also be presented with the results of eyewitness identification tests conducted by police. When considering evidence, jurors must evaluate the evidence—and the reliability of those presenting it—and come to a conclusion about the likely guilt of the defendant.

Given the important role that the courtroom plays in our legal system, it is not surprising that the various elements of court cases have prompted a considerable amount of psychological research. Court-relevant research first came to prominence with the publication of Munsterberg’s *On the Witness Stand* (1908). Some important research was conducted over the following half century, before the work of Strodtbeck and colleagues (e.g., Strodtbeck, James, & Hawkins, 1957) precipitated a surge in research on juror decisions. Loftus’ (e.g., Loftus, 1979) work on eyewitness memory represented another landmark in court-relevant research, and attention to eyewitness research within the legal system was increased further in the 1990s, when advances in DNA testing procedures—resulting in the exonerations of numerous innocent suspects—highlighted mistaken eyewitness identification as by far the leading cause of wrongful convictions (Innocence Project, 2010). At present, one of the most important challenges for researchers conducting court-relevant work is to improve communication with legal professionals so that research findings can translate to appropriate policies (Wells, 2005; Wells, et al., 2000).

In this chapter, we discuss the methodologies used to conduct research in three prominent areas of relevance to the court: eyewitness identification tests, interviewing of
Court relevant research methods

witnesses, and juror decision making. It is worth noting from the outset that some aspects of research design are the same across the areas. For example, experimental designs are the norm in each. Other aspects differ between areas. For example, the type of stimuli typically used in eyewitness identification studies differs from that used in juror studies. Such similarities and differences will be highlighted throughout the chapter.

Eyewitness Identification

Research Aims

Mistaken eyewitness identification is the leading contributor to wrongful convictions (Scheck, Neufeld, & Dwyer, 2003). To date (16th August 2010), 258 falsely imprisoned men in the USA have been exonerated through DNA testing by the Innocence Project (www.innocenceproject.org). Mistaken eyewitness identification played a central role in 75% of the first 239 cases. These eyewitnesses were not deliberately attempting to subvert the course of justice. They were using their memories of the culprit's face and appearance to make the best identification decision that they could. And while in many cases, eyewitnesses are able to correctly identify the culprit, there is clearly a large margin for error.

Psychologists’ aims in this area have been two-fold: protecting innocent suspects from mistaken identification, and increasing the chances that guilty suspects will be identified. The first aim has received the most attention, due to the dire consequences of mistaken identification. However, psychologists must ensure that any recommendations that they make to policy makers will not increase the likelihood of guilty suspects going free.

Before beginning an identification experiment, the researcher must have a clear research question in mind. For example, does the presence of a weapon during a crime affect witnesses’ abilities to accurately identify the perpetrator? Most identification experiments follow the same basic procedure. First, the participants are shown a staged crime, which
could be live or videotaped. Some period of delay passes, after which the participants are given an identification test. The researcher must choose an appropriate population to sample (young adults, children, older adults, people with learning disabilities), and must develop appropriate stimulus materials. For example, two similar videotaped crimes might be produced – one featuring a gun, and another with no visible weapon. Below are summarised some of the main methodological choices that must be made by any researcher who wishes to run an eyewitness identification experiment. This is followed by a brief discussion of the strengths and limitations of this kind of research, and a consideration of future directions in the field.

**Methodology**

*The mock crime.* Most eyewitness identification studies begin with a mock crime. This could be a live event (Valentine, Darling, & Memon, 2007) or a videotaped event (Semmler, Brewer, & Wells, 2004). Videotaped events allow greater experimental control, while sacrificing some external validity. Mock crimes may not always be practical: field studies may use mundane encounters between the participants and a confederate (Carlucci, Kieckhaeffer, Schwartz, Villalba, & Wright, in press; Wright, Boyd, & Tredoux, 2001). The content of the mock crime or encounter will be guided by the research questions, constraints on resources, and ethical considerations. For example, showing a violent mock crime to a sample of children would be inappropriate.

Participants are shown this mock crime. They may view the crime individually on videotape. Alternatively, researchers interested in social influences on memory may show participants the crime in pairs or small groups. Some studies have even involved live events occurring in lecture theatres to groups of several hundred participants (e.g. Malpass & Devine, 1981a). Participants are generally not told that they will need to identify the
perpetrator/confederate later. Rather, the mock crime is usually presented under some other guise (e.g. a study on memory for events).

*Retention interval.* Next follows some period of delay, which can range from minutes (Palmer, Brewer, McKinnon, & Weber, 2010) to days (Lefebvre, Marchand, Smith, & Connolly, 2007), to weeks or even months (Malpass & Devine, 1981a). The delay period will largely be guided by practical considerations, and may require adjustment if the task is too difficult or too easy for the participants.

*Selecting foils and measuring lineup fairness.* To create a line-up, the researcher must select some appropriate distractor faces, or “foils”. In a fair line-up, the foils should resemble the suspect. Foils can be selected in two ways: based on the physical resemblance to the culprit, or based on a witness’s description of the culprit. Luus and Wells (1991) argued that investigators should match-to-description rather than matching-to-appearance. However, this is usually impractical for a researcher, as a new line-up would be required for each participant witness who potentially could provide a different description. With large sample sizes required in eyewitness identification studies (see, for example, Palmer et al., 2010), this would be very costly and time-consuming. Most researchers, therefore, select foils to match the culprit's appearance. Others opt for a middle-ground approach, selecting foils based on descriptions from one or more independent participants.

Line-up fairness is as important in the lab as in the justice system. In the real world, unfair line-ups can lead to wrongful convictions; in the laboratory, unfair line-ups can create spurious results. Lineup fairness is evaluated using the mock witness paradigm. “Mock witnesses” read a brief description of the suspect, and are asked to select the suspect from the lineup. In a fair lineup, the choices should be evenly distributed across all of the line-up members. In a very unfair line-up (e.g. an African-American suspect with White foils), all of the mock witnesses would choose the suspect. There are several statistical techniques for
estimating line-up fairness, which make different assumptions (see Brigham, Meissner, & Wasserman, 1999, for a review). A simple measure is *functional size*, calculated by dividing the total number of mock witnesses by the number of mock witnesses who chose the suspect (Wells, Lieppe, & Ostrom, 1979). For example, if there are one hundred mock witnesses, and 50 choose the suspect, then the functional size of the lineup is 2 (100/50). This means that there is one plausible foil in addition to the suspect. If the functional size is close to the actual lineup size, then the lineup may be fair (though there are other statistical considerations to be made. See Tredoux, 1998, for a detailed review of measures of lineup fairness).

*Lineup presentation.* Line-up presentation can vary in modality - for example, photographic (Gorenstein & Ellsworth, 1980), video (Valentine et al., 2007), or live presentation (Malpass & Devine, 1981b). Photographic line-ups are easy to create, and are used by many police forces around the world. They are therefore favoured by researchers. However, a few police forces now use video line-ups as standard, and so some researchers have followed suit.

Researchers must decide whether to use a *simultaneous* or *sequential* line-up (Lindsay & Wells, 1985). In a simultaneous lineup, all of the images are seen together. In a sequential lineup, each image is seen individually and the participant makes a yes/no decision for each image. In some versions of the sequential lineup the line-up ends as soon as a positive identification is made. Sequential line-ups tend to produce higher accuracy than simultaneous line-ups. Sequential line-ups encourage absolute judgments ("Does this person match my memory of the culprit?") whereas simultaneous line-ups encourage relative judgments ("Which person is the best match to my memory of the culprit?"). Some studies use simultaneous line-ups (e.g. Valentine & Mesout, 2009) and others use sequential line-ups (e.g. Sauer, Brewer, & Wells, 2008). This choice will often be guided by the research questions and by practical constraints.
Target present and target absent lineups. Crucially, eyewitness identification studies should include both target present (TP) and target absent (TA) line-ups. TP line-ups contain the actual culprit from the crime event, while TA line-ups do not. In TA line-ups, the culprit can be replaced with a designated innocent suspect, chosen for high similarity to the culprit (e.g. Carlson, Gronlund, & Clark, 2008). Alternatively, the culprit can be removed without replacement. TA line-ups act as a vital control, allowing researchers to examine the behaviour of witnesses when the suspect is not the culprit. This is essential for making recommendations to policy makers and for drawing conclusions about eyewitness misidentification in the real world.

Instructions to witnesses. The instructions given to participants can dramatically influence choosing behaviour. Unbiased instructions, which state that the culprit may or may not be present, reduce inaccurate identifications of innocent suspects and foils (Malpass & Devine, 1981b). Many police forces are required to give unbiased instructions to witnesses, so researchers often use unbiased instructions in the lab. However, there are situations in which researchers will not use unbiased instructions, such as when the research question requires high choosing rates, and high error rates (e.g. Brewer, Weber, Clark, & Wells, 2008). Caution should then be exercised when applying the results to real world eyewitness behaviour.

Analysis

Participants' responses can be divided into several categories, shown in Table 1. A participant can make one of three responses. They can identify the suspect, they can identify a foil, or they can choose to reject the lineup by not making an identification. In both TP and TA lineups, foil IDs will always be inaccurate. Note, however, that in TP line-ups, the correct response is to identify the suspect, but in the TA line-up, the correct response is to reject the line-up (correct responses shown in italics). Also note that incorrect suspect IDs can only be
observed when an innocent suspect was designated. Without an innocent suspect, this data cell will be empty.

Table 1. *Response categories for target present and target absent lineups. Correct responses are shown in italics.*

<table>
<thead>
<tr>
<th>Target present line-up</th>
<th>Correct suspect ID</th>
<th>Foil ID</th>
<th>Lineup rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target absent line-up</td>
<td>Incorrect suspect ID</td>
<td>Foil ID</td>
<td>Lineup rejection</td>
</tr>
</tbody>
</table>

The frequencies of the participants’ responses are recorded. As these data are categorical, they should be analysed with $\chi^2$ tests or log linear analysis, depending on the research design and questions (see Field, 2009, for an excellent introductory book on statistics for psychology). The observed frequencies are compared against some expected frequencies - for example, chance performance. The analyses will tell the researcher whether the observed results significantly differ from the expected results.

**Worked Example**

The following example is taken from Malpass and Devine (1981b). Note that some methodological details have been omitted here for brevity; a comprehensive description can be found in the original article. Malpass and Devine were interested in whether the instructions given to eyewitnesses would influence the accuracy of identification decisions. Specifically, they asked whether the rate of false identifications from lineups could be reduced simply by reminding witnesses that the person they were looking for might not be in the lineup. To address this question, they used a 2 (lineup instructions: biased or unbiased) $\times$ 2 (lineup type: target-present or target-absent) between-subjects experimental design. The four conditions are shown in Table 2.
Table 2. *The four experimental conditions created by a 2 x 2 between-subjects design.*

<table>
<thead>
<tr>
<th>Unbiased instructions</th>
<th>Unbiased instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target-present</td>
<td>Target-absent</td>
</tr>
<tr>
<td>Biased instructions</td>
<td>Biased instructions</td>
</tr>
<tr>
<td>Target-present</td>
<td>Target-absent</td>
</tr>
</tbody>
</table>

Malpass and Devine (1981b) had 100 participants view a staged mock-criminal event. The crime took place in a classroom setting and involved a male culprit (actually a confederate of the experimenters) interrupting a class presentation, arguing with the presenter, and pushing over some EEG equipment before fleeing the scene. The participants were then informed that the crime had been staged and, later, were asked to attempt to identify the culprit from a lineup.

Participants viewed either a target-present or target-absent lineup. For the target-present lineup, the culprit appeared alongside four other males of similar height, build, hair colour and style. For the target-absent lineup, the culprit was replaced by another appropriate foil. None of the lineup members wore clothing similar to that worn by the culprit during the staged crime. The position of lineup members was rotated so that each person appeared in each position equally often. Live presentation was used, with the lineup members located approximately 7 metres from witnesses, who viewed the lineup through a 2-way mirror.

Participants were randomly allocated to receive either biased or unbiased lineup instructions. In both conditions, instructions were standardised by having the lineup administrator read them aloud and participants indicated their identification decision on a response sheet. The biased instructions read: “We believe that the person who pushed over the electronics equipment during the EEG demonstration is present in the lineup. Look carefully at each of the five individuals in the lineup. Which of these is the person you saw
push over the equipment? Circle the number of his position in the lineup below.” (Malpass & Devine, 1981b, p. 484.) The unbiased instructions read: “The person who pushed over the electronics equipment during the EEG demonstration may be one of the five individuals in the lineup. Look carefully at each of the five individuals in the lineup. If the person you saw push over the equipment is not in the lineup, circle 0. If the person is present in the lineup, circle the number of his position.” (Malpass & Devine, p. 484).

Analyses were performed using a variation on the chi-square test (see Langer & Abelson, 1972). The results indicated that lineup instructions had a substantial effect on identification accuracy. When the culprit was not in the lineup, participants made more foil identifications if they received biased versus unbiased instructions. Importantly however, when the culprit was in the lineup, the rate of correct identifications was not lower for unbiased instructions (83%) than biased instructions (75%). We now know that the use of unbiased instructions can reduce the occurrence of false identifications without markedly reducing the rate of correct identifications, thus improving overall accuracy.

**Strengths and Limitations**

Laboratory studies allow us to control factors that cannot be controlled in the real world. To understand the effects of delay on eyewitness identification, we can hold all other factors constant, randomly allocating participants to different delay conditions. Individual studies will always be limited in their conclusions. However, psychologists have now created a rich empirical literature which has provoked real change in police practice (see Wells et al., 2000).

Of course, there are many ways in which experiments differ from crimes witnessed by real witnesses. These differences have prompted concerns about low ‘external validity’. That is, the worry that findings from the lab will be not extend to real world behavior. For example, lab studies rely on videotaped mock crimes, or on fairly mundane live events,
perhaps involving theft of equipment from the lab (Carlson et al., 2008, Experiment 1). It would be unethical to show participants deeply distressing violent events, or to make them feel personally endangered. However, some creative researchers have taken advantage of naturally occurring high-stress situations to examine eyewitness memory. For example, Morgan and colleagues (2004) tested participants’ memories for an interrogator from a high-stress interview which had formed part of a military training program. Valentine and Mesout (2009) asked participants to identify a confederate who had approached them while walking through a section of the London Dungeon which was created to elevate stress and fear responses. Such research bridges the gap between the experience of participants in the lab and witnesses in the real world.

A constant problem in eyewitness identification research is statistical power (see Cohen, 1992, for a discussion of statistical power). Eyewitness experiments often involve designs which require each participant to be allocated to a different group. Usually, each participant will only make one identification decision from a single lineup. As a result of this, required sample sizes can reach several hundred (see Palmer et al., 2010; Sauer et al., 2008), making research time consuming and costly. Where possible, researchers can design experiments so that each witness sees more than one line-up to increase statistical power.

A crucial concept within experimental science is “generalisability”. Researchers do not want to limit their conclusions to the specific set of conditions within a given experiment. Rather, they wish to extend their results to a more general set of circumstances. In the eyewitness domain, the generalisability of a study’s results can be influenced by many factors, including the distinctiveness of the target. If the culprit is especially distinctive, or not at all distinctive, the results may not be replicable in other situations (Wells & Windschitl, 1999). These concerns are compounded when the same small set of stimulus materials are shared among different researchers. One often used remedy is to use two or more culprits,
counterbalanced between participants, thereby permitting an examination of whether findings are robust across a wide variety of stimulus materials and conditions.

**Future Directions**

Our understanding of memory has changed dramatically over the last few decades. Long gone are the days in which we believed that the brain was a video camera, able to perfectly record visual input which was neatly filed away for later viewing. We now know that memory is malleable and fragile, and that people make mistakes even when viewing conditions are optimal. Through communicating these ideas to policy makers, eyewitness identification procedures have improved (see Wells et al., 2000).

However, there are limits to how much the basic line-up procedure can be improved. The UK uses some of the most advanced identification procedures in the world (see Kemp, Pike, & Brace, 2001). Yet archival analyses of real line-up outcomes in the UK show that suspect identifications remain low (41% - 44%, Memon, Havard, Clifford, Gabbert, & Watt, in press). Perhaps it is time to rethink the line-up entirely, and to come up with radically different alternatives which are grounded in theory (Wells, Memon, & Penrod, 2006). For example, Sauer, Brewer, and Weber (2008) tested a procedure where, rather than being asked to make an absolute decision about whether the culprit is in the line-up, the participants rated how confident they were that each line-up member was the culprit. These confidence ratings were better able to predict the identity of the culprit than were standard binary identification decisions. This program of research, still in its early stages, offers a promising new avenue of investigation for researchers interested in eyewitness memory.

The traditional line-up pre-dates any real scientific understanding of human memory (Davies & Griffiths, 2008). Psychologists have focused on improving this procedure because of its wide usage. However, we have a vast body of knowledge on human perception and memory which we can use to develop a new paradigm for eyewitness identification. It is
possible that any such advance would be likely resisted by the courts. However, if the science is conducted rigidly, and the weight of evidence suggests a dramatic improvement in eyewitness accuracy, then an overhaul of identification procedures may be possible (see, for example, Brewer & Weber, 2008; Sauer et al., 2008).

Interviewing Witnesses

Research Aims

An eyewitness’s memory of an event is often crucial for investigators. But memory is fragile and vulnerable, prone to distortion and deterioration. Early research exposed the fallibility of eyewitnesses and their susceptibility to misinformation (e.g. Loftus & Palmer, 1974). Researchers then began developing interview techniques to maximise the amount of accurate information gained from witnesses without increasing inaccurate information. The most successful technique was the cognitive interview (CI), an interviewing framework grounded in psychological theory on memory encoding, storage, and retrieval (Geiselman et al., 1984). The original CI had four components:

i) Context reinstatement. The witness thinks about the physical environment and the personal context (thoughts and feelings) of the event. For example, a witness to a car accident would imagine themselves at the scene of the accident. They would be asked to recall physical aspects of the scene (where they were located in relation to other parts of the environment, weather conditions, sounds that they heard etc.) as well as the thoughts and emotions that they had prior to and immediately after the accident.

ii) Report everything. The witness reports everything that he or she can remember, without interruption. For example, the witness would tell the interviewer about the accident, in as much detail as possible. It is very
important that the interviewer allows the witness to use their own words, and that this free recall is not interrupted.

iii) Change temporal order. The witness recalls events in a different chronological order, hopefully cueing additional details. For example, the interviewer could ask the witness to begin by describing what happened after the accident, then work their way back from that point.

iv) Change perspectives. The witness recalls the event from different perspectives, potentially providing new memory cues. For example, the witness could imagine what the accident would have looked like from the opposite side of the street, or from above.

The CI has been modified many times, often by omitting the ‘change perspectives’ component, and adding in rapport building (e.g. Dando, Wilcock, Milne, & Henry, 2009). This is when the interviewer establishes a rapport with the witness, putting them at ease and explaining the interview procedure to them. The original CI and its modifications are described in detail elsewhere (see Memon, Meissner, & Fraser, in press). This section will outline the major methodological considerations for eyewitness interviewing research. Strengths and limitations will be discussed, and future directions will be considered.

Methodology

Interview type. Researchers usually use the CI, often modified from its original format. The enhanced CI, for example, includes elements such as rapport building and transfer of control to the witness, where the interviewer makes it clear to the witness that he or she is in control of the interview, and that he or she has the knowledge that the interviewer is hoping to access (Fisher & Geiselman, 1992). If working with special populations (e.g., children, older adults, cognitively impaired), researchers may omit some components of the CI, such as the change temporal order and change perspectives components, as these are more
cognitively demanding than the context reinstatement and report everything components and may be difficult for such populations to understand (e.g., Wright & Holliday, 2007). Furthermore, the context reinstatement and report everything components together have been found to be more effective than any of the other components, even when interviewing children as young as 5 years old (Milne & Bull, 2002).

**Control groups.** To draw conclusions about an interview technique, there must be a control group. Control groups are essential for any experimental study, as they allow the comparison of one group who received the experimental treatment (the CI) with another group who did not receive the experimental treatment. If all other variables are kept equal, and if the participants are randomly allocated to the groups, then the researcher can infer that any differences between the groups are due to the treatment (the CI). The control group usually receives a structured interview (SI), which shares some common features with the CI – rapport building, free recall, follow-up questions related to the content of the free recall. However, a SI excludes many features of the CI, including context reinstatement, ‘report everything’ instructions, and ‘change in temporal order’/’change perspectives’ (e.g., Wright & Holliday, 2007). The components featured in SIs and CIs are compared in Table 3.

**Coding and scoring.** Before accuracy can be assessed, a comprehensive list of all of the details in the crime event must be made. This must be done carefully, and the responses of several researchers can be pooled together (see Davis, McMahon, & Greenwood, 2005). This list then serves as a template against which the interviews can be scored.

Table 3. *Interview components featured in structured interviews and cognitive interviews.*

<table>
<thead>
<tr>
<th><strong>Structured interview</strong></th>
<th><strong>Cognitive interview</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapport building</td>
<td>Rapport building</td>
</tr>
<tr>
<td>Transfer of control to the witness</td>
<td>Transfer of control to the witness</td>
</tr>
<tr>
<td>---</td>
<td>Context reinstatement</td>
</tr>
</tbody>
</table>
Each detail mentioned within an interview is scored (with each unique detail scored only once). Commonly used scoring categories include: correct details, errors (recalling a red hat instead of a green hat), confabulations (details which were not present), and intrusions (details which came from some other source, such as a leading question; see Memon, Holley, Wark, Bull, & Köhnken, 1996). Details may also be coded by content, such as person, action, and object details (e.g. Akehurst, Milne, & Köhnken, 2003). Two or more naïve coders should code the interviews separately. Inter-rater reliability is then checked using statistical indicators of agreement, such as Cohen’s kappa (see Field, 2009). Any coding discrepancies are resolved, usually through discussion between the coders. If any disagreements remain, these can be resolved by an independent third party.

**Analysis**

For each witness, the number of details in each response category is summed. These scores can be compared using a range of statistical methods, including t tests, ANOVAs, and regressions (Field, 2009). The statistics will depend upon the experimental design, the number of experimental groups, and the hypotheses.

**Worked Example**

The following example is taken from Larsson, Granhag, and Spjut (2003). Again, some of the methodological and analytical details have been omitted here for brevity. Larsson et al. tested children aged 10-11 years old either one week or six months following a
videotaped event. The children were randomly allocated to receive a CI or a SI. The experiment therefore had a 2 x 2 between subjects design. The four experimental conditions are summarised in Table 4.

Table 4. The four experimental conditions from Larsson et al. (2003).

<table>
<thead>
<tr>
<th>One week delay</th>
<th>Six month delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Interview (CI)</td>
<td>Cognitive Interview (CI)</td>
</tr>
<tr>
<td>One week delay</td>
<td>Six month delay</td>
</tr>
<tr>
<td>Structured interview (SI)</td>
<td>Structured interview (SI)</td>
</tr>
</tbody>
</table>

Forty nine participants took part, recruited from a local school. All of the children watched a videotaped event, in which a fakir performed several amazing feats, including playing with fire and laying on a bed of nails. The children were instructed to pay close attention to the video, but were not told that they would receive an interview on the contents of the video.

Following other researchers, the change temporal order and change perspectives components were excluded from the CI, to make it more suitable for use with children. The context reinstatement and report everything components were included, and these components were the key differences between the CI and the SI. After a delay of one week, one quarter of the original sample were given a CI and one quarter were given a SI. The remaining children were interviewed with either a CI or a SI after six months. The interviews were conducted by three interviewers, all of whom received training in the CI, and all of whom had previous experience in interviewing children.

A checklist of all of the information in the film was prepared. The checklist contained 224 distinct units of information. Two raters then coded the children’s interviews, recording
each unique detail mentioned. These details were then compared to the checklist to provide
the following measures: correct information, incorrect information, and confabulations
(details which were not in the video). Percentage accuracy and completeness of the reports
were then calculated for each participant. One rater coded all of the interviews, and a subset
were randomly selected to be coded by a second rater. Inter-rater agreement was high (93%),
and any discrepancies were resolved by discussion.

Larsson et al. (2003) compared recall performance in the four conditions with a
series of between subjects ANOVAs (one for each measure). They found that children who
received the CI recalled more correct information and gave more complete and accurate
accounts than children who received the SI. Children interviewed after six months recalled
less correct information and reported fewer confabulations than children interviewed after
one week. Longer delays led to lower overall accuracy and less complete accounts. Interview
type and delay did not interact on any measure – that is, the effect of delay was not
influenced by the effects of interview type. The authors concluded that the (modified) CI is a
useful tool for interviewing children, even after long delay periods.

Strengths and Limitations

Interviewing research shares many limitations with identification research. Reliance
on videotaped mock crimes or low arousal staged events still raises concerns of low external
validity. Specifics of the mock crime which is used could also create results which do not
generalize to other stimulus materials. In addition, researchers must ensure that the standard
of the interviews is high, and that the interview style is consistent if multiple interviewers are
used. Interviewers should therefore be trained in the CI. This can create some practical
difficulties; training can be expensive and time consuming to run, and the research team must
gain access to appropriate training materials. However, despite these difficulties, proper
training in CI techniques is essential for any group wishing to research eyewitness memory
using the CI, to ensure that their procedures conform to the recognized standard. The exact interview procedure used should be reported explicitly when publishing, as there are so many modified versions of the CI in the literature. So, for example, details of any components which are included (e.g. context reinstatement, report everything), and details of any components which are excluded (e.g. change temporal order, change perspectives) must be presented for publication.

Memon et al. (in press) discuss some of the gaps in our current knowledge of interviewing, such as how real frontline investigators use the CI. Most studies use student or academic interviewers, with only a small minority recruiting trained law enforcement professionals as interviewers (see Geiselman, Fisher, MacKinnon, & Holland, 1985). There are practical reasons for this; law enforcement agencies can be difficult to recruit into research programs due to high demands on their time, and already over-stretched budgets. However, to convince policy makers, more researchers should recruit police interviewers in eyewitness research.

**Future Directions**

A promising new line of research is the grain-size interview (Weber & Brewer, 2008), which builds on a theoretical framework of memory reporting. Goldsmith, Koriat, and Weinberg-Eliezer (2002) noted that information varies in ‘grain size’, from coarse (“He was wearing a dark jacket”) to fine grained (“He was wearing a navy blue jacket with a white logo on the right sleeve”). When reporting details, the individual controls the grain size of the information, aiming for an optimum balance between informativeness and accuracy. Weber and Brewer asked participants to provide both a fine grained response and a coarse response to a series of questions about a mock crime. The participants rated their confidence in each answer, and chose which they would like to volunteer as their preferred answer. Fine grained answers were chosen 35% of the time, and were volunteered when the benefits of
informativeness were likely to be large, and the costs to accuracy were likely to be small.

While still in its infancy, grain size interviewing offers interesting possibilities for
maximising information gain from witnesses. For example, preliminary research suggests that
the technique has the capacity to elicit a large amount of accurate coarse grained information
that witnesses would not otherwise report (Brewer, Muller, Nagesh, Hope, & Gabbert, 2010;
Hope, Gabbert, Brewer, Tull, & Nagesh, 2010).

Juror Decision Making

Research Aims

In this section, we concentrate mainly on decisions made by individual jurors (cf. collective verdicts arrived at by juries after deliberation), which have been the focus of the majority of research in this area. The study of juror decisions is concerned with several prominent questions, such as: What are the factors that influence jurors’ decisions? How do jurors process information during a trial? How well do jurors comprehend judicial instructions about points of law? And, are jurors competent decision makers? Approaches for addressing such questions fall broadly into two categories: field studies and mock-juror experiments. Below, we discuss some key methodological aspects of each.

Methodology

Field studies. Field studies deal with data from actual trials and, hence, are excellent in terms of realism. However, field studies give researchers little or no experimental control over variables of interest. For example, a researcher cannot, practically or ethically, arrange for a particular piece of evidence to be given to one randomly-selected half of a jury but not
the other half. This problem limits the extent to which causal conclusions can be drawn from field studies.

In addition, because jurors are typically not allowed to discuss their opinions during an active case with anyone—including researchers—field studies often involve collecting data via post-trial interviews or surveys of jurors. One drawback of this approach is that it relies heavily on jurors’ memories. To solve this problem, some studies (e.g., McCabe & Purves, 1974) have employed a *shadow jury* paradigm, whereby a second jury is selected by the researcher (ideally from the same pool that the actual jury was drawn from) to sit in the gallery of the court and listen to the case. The researcher is then able to obtain data from shadow jurors during the course of the trial.

*Mock-juror experiments.* Mock-juror experiments deal with decisions made by participants (typically jury-eligible individuals) who are presented with evidence and asked to behave as if they are jurors in a real case. The mock-juror paradigm allows researchers to experimentally manipulate a wide range of variables, some examples of which are discussed below.

*Presentation mode.* The degree of realism involved in mock-juror experiments varies widely, from having live actors present evidence in a real courtroom setting (e.g., Borgida, DeBono, & Buckman, 1990) to having student participants reading relatively brief portions of written transcript from a fictitious trial (for reviews, see Bornstein, 1999; Bray & Kerr, 1979). Constraints on time and financial resources usually dictate the choice of presentation mode.

*Defendant characteristics.* Juror decisions have been shown to be affected by certain characteristics of the defendant, such as attractiveness (e.g., Baumeister & Darley, 1982) and race (e.g., Johnson, Whitestone, Jackson, & Gatto, 1995). Apart from investigating the direct effects of defendant characteristics on juror decisions, researchers may wish to consider whether the effects of other variables are moderated by defendant characteristics.
Evidentiary variables. Researchers have innumerable options when it comes to manipulating aspects of the evidence presented to mock-jurors. A few examples of evidentiary factors include the complexity of evidence (e.g., Heuer & Penrod, 1994), the number of inconsistencies in a witness’s testimony (e.g., Berman & Cutler, 1996), the presence of testimony from expert witnesses (e.g., McKimmie, Newton, Terry, & Schuller, 2004), and the confidence with which witnesses deliver testimony (e.g., Brewer & Burke, 2002). Importantly, the choice of variables to be manipulated—and specific details of the manipulations—should be guided by the research questions being addressed.

Collective juror decisions. Although the majority of research has examined decisions made by individual jurors, collective jury decisions (i.e., verdict delivered after deliberation) have also been investigated. Studies of collective juror decisions have taken numerous approaches. For instance, the consistency and competency of jury decisions has been examined by comparing the verdicts given by actual juries with those given by mock juries (McCabe & Purves, 1974) and presiding judges (Kalven & Zeisel, 1966). Other researchers have investigated the occurrence of social psychological phenomena in the context of jury decisions. For example, several mock juror experiments (e.g., Bray & Noble, 1978; Hastie, Penrod, & Pennington, 1983) have found evidence of group polarization, whereby the views held by jurors tend to become stronger and more extreme as the jury interacts. Although not widely used, one interesting option open to researchers in this area is to use confederates who pose as participants but have been asked to act in a certain pre-determined manner during jury deliberations (e.g., to steadfastly oppose the majority opinion). The use of confederates affords researchers experimental control over a wider range of variables that come into play during the deliberation stage of juror decisions.

Judicial instructions. One important area of research addresses the impact of judicial instructions (i.e., about relevant points of law) on juror decisions. Such research can be
conducted via field studies or mock-juror experiments. Some issues examined include the extent to which jurors comprehend judicial instructions (Ogloff & Rose, 2005), and whether comprehension is affected by the timing (ForsterLee, Horowitz, & Bourgeois, 1993) and mode of presentation of instructions (Brewer, Harvey, & Semmler, 2004).

Deception detection. Because witnesses are not always co-operative, jurors may sometimes attempt to make judgments about truthfulness based on verbal or non-verbal cues exhibited by a witness. Similarly, jurors in some countries may be presented with results from physiological lie-detection tests as evidence that bears on the truthfulness of a witness’s account. Readers who are interested in conducting research in this area (e.g., examining the ability of jurors—or police investigators—to detect deception or interpret polygraph results) are encouraged to refer to Vrij (2008) for an excellent coverage of the extensive deception detection literature, which will guide the selection of research questions and variables to be manipulated.

Manipulation checks. In many mock-juror experiments, it is appropriate for researchers to include checks to test whether their experimental manipulations have been successful. Manipulation checks are often straightforward. For example, a manipulation of testimonial consistency (high vs. low) might be assessed by asking participants to recall the number of inconsistencies that occurred in the evidence they were presented with (as per Brewer & Burke, 2002). If the manipulation was successful, participants in the high consistency condition should, on average, recall fewer inconsistencies than those in the low consistency condition. To reduce the chances of participants becoming aware of the purpose of a study, researchers should have participants complete any manipulation checks after the more important measures have been obtained. For example, mock-jurors might be asked to recall the number of inconsistencies in a witness’s testimony (i.e., the manipulation check)
after rating the credibility of the witness and giving a guilty/not guilty verdict (i.e., the measures of greatest interest to the researcher).

**Analyses**

For field studies and mock-juror experiments, the most common dependent measures are dichotomous verdicts (i.e., guilty or not guilty) and continuous ratings of probable guilt (e.g., from 0% - *definitely not guilty* to 100% - *definitely guilty*). Analyses of verdict data involve chi-square tests and hierarchical loglinear analysis, while analyses of probable guilt ratings involve t-tests and factorial ANOVA. One important point to note is that, in field studies, variables that are presumed to be causal must be measured rather than manipulated. For example, Heuer and Penrod (1994) were interested in the effects of trial complexity on jury decisions in actual cases. The researchers obviously had no control over the complexity of actual cases and, therefore, had to measure complexity instead of manipulating it. They did this by asking the trial judge in each case to rate the complexity of that case—compared to an average case—in terms of a variety of aspects (e.g., the evidence presented, the relevant points of law, and the arguments made by the prosecution and defense counsels).

**Worked Example**

The following example is taken from Brewer and Burke (2002), who investigated the effects of testimonial inconsistency and eyewitness confidence on verdicts in a mock-juror experiment. (Again, some methodological details have been omitted here for brevity; a comprehensive description appears in the original article.) Their focus was on the interaction between the two independent variables. For example, would the presence of inconsistencies affect juror judgments regardless of witness confidence, as suggested by surveys of police, lawyers, and jury-eligible individuals, in which inconsistencies are rated as a strong indicator of inaccuracy (e.g., Brewer, Potter, Fisher, Bond, & Luszcz, 1999)? Alternatively, would inconsistencies affect juror decisions only when witness confidence was low, as implied by
the results of Lindsay, Wells, and Rumpel (1981)? Or would confidence “trump” inconsistency, as suggested by the results of Cutler, Penrod, and Stuve (1988)?

The stimulus materials (based on materials developed by Berman and Cutler, 1996) comprised a 20-minute, audio-taped transcript of a prosecution witness being questioned by prosecution and defense attorneys during a bank robbery trial. Brewer and Burke (2002) developed four versions of the materials, one for each experimental condition in their 2 (testimonial consistency: consistent or inconsistent) × 2 (witness confidence: high or low) design. Each participant heard one version of the transcript: a high-confidence witness giving consistent testimony, a high-confidence witness giving inconsistent testimony, a low-confidence witness giving consistent testimony, or a low-confidence witness giving inconsistent testimony.

For the consistency manipulation, the critical items were the witness’s answers to four questions that were asked twice, first by the prosecution attorney and then by the defense attorney during cross-examination. In the consistent testimony condition, the witness’s responses to these four questions did not differ between the first and second time they were asked. In contrast, in the inconsistent condition, the witness’s responses under cross-examination clearly contradicted their earlier responses (e.g., the robber threw the money into a canvas bag vs. the robber put the money in his pocket). These differences were highlighted by the defense attorney.

Confidence was also manipulated via the witness’s responses to questions. In the high confidence condition, the witness answered all questions without hesitation or qualification of her responses (e.g., “Yes, exactly”; “No”). In contrast, in the low confidence condition, the witness did hesitate and qualify when answering some questions (e.g., “Yes, I guess so”; “Ummm…I don’t think so”).
Brewer and Burke (2002) recruited 130 jury-eligible participants. Each participant was asked to imagine that they were a juror in an actual case, and was randomly assigned to hear one of the four versions of the transcript. Participants were then asked to indicate the probability that the defendant committed the crime (on a scale from 0% - *I am not sure at all that the defendant committed the crime* to 100% - *I am 100% sure that the defendant committed the crime*), and to give a guilty/not guilty verdict. As manipulation checks, participants were then asked to rate the confidence with which the defendant delivered her testimony (on a scale from 1 – *not very confidently* to 7 - *extremely confidently*) and to recall how many, if any, contradictions occurred in the testimony. These items were embedded among several questions unrelated to the purpose of the experiment. The results of the manipulation checks suggested that the manipulations had the desired effects: On average, participants in the high confidence condition rated the witness as more confident than those in the low confidence condition (*M* = 5.98 vs. 3.02), and participants in the inconsistent condition recalled more contradictions than those in the consistent condition (*M* = 3.31 vs. 0.94).

The results for the central dependent measures suggested that, in line with the results of Cutler et al. (1988), participants’ decisions were influenced primarily by witness confidence. A 2 (witness confidence) × 2 (testimonial consistency) loglinear analysis of verdicts indicated that participants gave more guilty verdicts when the prosecution witness was high in confidence (39.4% guilty) versus low in confidence (9.4% guilty), and this pattern held regardless of whether the testimony was consistent or inconsistent. Similarly, a 2 (witness confidence) × 2 (testimonial consistency) factorial ANOVA showed that participants gave higher probable guilt ratings when the witness was high (57.50%) rather than low (32.50%) in confidence, and this pattern occurred in both testimonial consistency conditions. (See Brewer and Burke, 2002, for a detailed interpretation of these results.)
Strengths and Limitations

Compared to mock-juror experiments, field studies have a clear advantage in realism. For example, the decisions made by jurors in actual cases carry real consequences, while those made by mock-jurors and shadow jurors do not. The lack of realism in many mock-juror experiments has given rise to a considerable amount of criticism (e.g., Weiten & Diamond, 1979). However, there is evidence suggesting that at least some of this criticism is not valid. For example, some studies have found no meaningful difference in the verdicts given by jurors who anticipate that their decisions will carry real consequences for defendants, and jurors who do not (Kerr, Nerenz, & Herrick, 1979). In addition, as outlined earlier, one very important disadvantage of field studies is that they do not afford researchers experimental control over variables, which limits the extent to which causal conclusions can be drawn. In contrast, well-conducted mock-juror experiments do enable researchers to draw causal conclusions.

Future Directions

Given the methodological and practical advantages of mock-juror experiments, it is important for future research to further investigate the extent to which results found using such methods can be generalized to more realistic jury settings. Some researchers (e.g., Kerr & Bray, 2005) have advocated a compromise, whereby psycho-legal research is conducted using a variety of field study and experimental methods. In this vein, researchers might begin by attempting to answer a question using relatively simple, well-controlled experiments before trying to replicate the results in increasingly more realistic scenarios. This approach might be particularly fruitful for areas of juror decision making research that have the potential to guide policy, such as the development of procedures that lead to better processing of evidence or comprehension of law by jurors (e.g., Brewer, Harvey, & Semmler, 2004).

Further Reading
In this chapter, we have given an overview of the methodological questions that arise in three prominent areas of court-relevant research. However, these are not the only such areas. Productive research has been conducted on a wide range of topics that we have only considered briefly (e.g., deception detection; Vrij, 2008) or have not covered at all, such as the effects of pre-trial publicity on juror decisions (e.g., Studebaker, Robbennolt, Parthak-Sharma, & Penrod, 2000) and persuasive techniques used by attorneys during courtroom trials (e.g., Williams, Bourgeois, & Croyle, 1993). There are some excellent sources for readers who are interested in a broader and more detailed coverage of court-related research, including Brewer and Williams (2005), Hastie (1993), Hastie, Penrod, and Pennington (1983), Loftus (1979) and Vrij (2008). In addition, we highly recommend Pelham and Blanton (2007) as a useful starting point for any readers interested in learning more about the general principles of research methodology for psychology.
References


Author Note

Matthew A. Palmer, Ruth Horry, and Neil Brewer, School of Psychology, Flinders University, Adelaide, Australia.

This research was supported by Australian Research Council Grant DP36065.